ParagraphS
Paragraphs Environmental Design

Original Research

DOI: 10.5281/zenodo.24161079

OPEN ACCESS

***CORRESPONDENCE**
Matthew W. Fuchs, P.E.
fuchsw42@gmail.com

RECEIVED: June 2025
ACCEPTED: July 2025
PUBLISHED: September 2025

ISSN-L 2960-2874

VOLUME 3, ISSUE 3
2025

CITATION
Fuchs MW. (2025). Residential Density and Transit Proximity in Selected Metropolitan Counties: Evidence from the Smart Location Database. Paragraphs Environmental Design, 3(3), 40-59. DOI: 10.5281/zenodo.24161079.

COPYRIGHT
© 2025 Fuchs MW. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Residential Density and Transit Proximity in Selected Metropolitan Counties: Evidence from the Smart Location Database

Matthew W. Fuchs, P.E.*

Former chair of the Architectural Engineering and Building Construction Department,
Milwaukee School of Engineering

Abstract

Residential density is often treated as a foundation for effective public transport, yet proximity to a stop, scheduled service, and access to destinations measure different conditions. This article analyzes original estimates from the 2021 Smart Location Database for New York, Cook, Los Angeles, King, Harris, and Miami-Dade counties. The purposive sample contains 16,580 eligible residential block groups; 14,156 have usable transit-distance records. Descriptive comparisons, county fixed-effects models, tract-clustered uncertainty estimates, and sensitivity analyses distinguish recorded proximity from distance conditional on a valid observation. A doubling of residential density is associated with a 12.8% lower fitted geometric mean of distance plus one meter after adjustment for employment density, weighted intersection density, employment-household entropy, and county effects. The association persists under alternative weighting, density definitions, and exclusions. However, New York has universal recorded proximity within 800 meters and almost no within-county density-distance correlation, whereas Harris has a much larger unresolved share and substantially lower population-weighted recorded proximity. Logistic sensitivity analysis corroborates the positive density association with recorded 800-meter proximity in the other five counties. These findings support a contextual account of housing-transit alignment while rejecting causal interpretations and universal density thresholds. The contribution is a reproducible comparison that identifies where density corresponds with documented transit proximity and explains why data coverage, geographic scale, service conditions, and distribution must remain separate parts of its policy interpretation.

Keywords: residential density; transit proximity; Smart Location Database; metropolitan counties; accessibility; spatial analysis

10.5281/zenodo.24161079

Introduction

Residential concentration and public transport are often treated as mutually reinforcing components of metropolitan development. Concentrating homes within reach of transit appears to provide a practical basis for coordinating housing and mobility: more residents can potentially share an existing network, while nearby services can broaden the destinations available without driving. Yet this proposition combines several relationships that need not occur together. Dense housing may be distant from recorded stops, a nearby stop may offer infrequent service, and a well-connected neighborhood may remain inaccessible to households unable to afford its housing. The empirical starting point is therefore the correspondence between where housing is concentrated and where transit is recorded, rather than an assumption that residential density already constitutes effective transport access.

This correspondence matters because it can reveal whether the spatial distribution of housing supports or frustrates an important planning objective. A concentration of homes creates the possibility of serving many residents within a relatively small area, but the existence of potential demand does not itself specify an operating budget, route alignment, service span, or institutional responsibility. Equally, transit infrastructure can be present where residential development remains limited, whether because the surrounding land accommodates other activities or because housing construction faces constraints. The analytical implication is that density and proximity should be understood as distinct conditions whose alignment requires investigation. Their correlation can identify a pattern of coordination, while departures from that pattern can be equally informative about the questions that local investigation should pursue.

This article examines that narrower relationship in selected metropolitan counties using block-group observations from the Smart Location Database. The central question is whether denser residential areas are more closely associated with recorded transit access, and whether that association has a similar form across contrasting county contexts. A comparison of counties can make differences in urban structure visible, but counties are neither interchangeable metropolitan systems nor internally uniform neighborhoods. The analysis therefore treats them as explicitly selected comparative settings and retains smaller geographic units as the observations. This design permits scrutiny of both broad contrasts and internal variation, while avoiding an unsupported claim that the chosen counties represent the full distribution of metropolitan development in the United States.

The article's contribution lies in connecting a reproducible descriptive analysis with a disciplined interpretation of what its indicators can establish. Publicly available spatial data make comparison feasible, but apparent precision can conceal differences between recorded service availability, measured distance, and usable access. A meaningful result must therefore explain not only the direction of an association but also the population of places to which it applies and the observations that are excluded. The argument developed below is that density can be informative about the spatial organization of transit without becoming a sufficient explanation of that organization. This position gives the empirical analysis a clear purpose: to identify conditional patterns and mismatches that can guide further investigation, rather than to turn a cross-sectional coefficient into a universal planning rule.

The findings show a negative adjusted density-distance association alongside two different boundary cases: New York has universal recorded proximity within 800 meters and almost no internal distance gradient, whereas Harris has much less complete recorded proximity and a substantial population-weighting difference. Neither case supports a universal density rule. Together, they motivate an interpretation that separates the extent of documented proximity from differences among already covered neighborhoods.

Literature and conceptual framework

Density as an enabling condition

Residential density expresses a relationship between a quantity of housing and a quantity of land, making its denominator central to its meaning. The same number of dwellings can yield different density values depending on whether the

area includes extensive undeveloped land or is restricted to land classified as developed. Conversely, two areas with the same density can contain different dwelling types, occupancies, or distributions of homes within their boundaries. These differences do not invalidate density as a useful descriptor; they establish why its definition must remain consistent throughout a comparison. The present analysis consequently treats density as a measured intensity of residential development, while reserving claims about population exposure, household behavior, or development capacity for indicators that actually measure those concepts.

The 3Ds framework provides a reason to avoid interpreting residential intensity in isolation. Cervero and Kockelman (1997) separated density from diversity and pedestrian-oriented design, recognizing that these dimensions may jointly characterize a neighborhood while retaining distinct relationships with travel. That framework does not imply that each dimension can be changed independently in practice, nor does a statistical model automatically recover an isolated policy effect. Its implication for a density-proximity study is more limited and more useful: neighborhoods with similar residential intensity may have different street patterns or mixes of activity. A measured density association can therefore be an informative feature of a larger urban configuration without being evidence that density is the sole mechanism producing transit proximity.

The subsequent expansion to five commonly discussed dimensions strengthens this distinction. Ewing and Cervero's (2010) meta-analysis considered destination accessibility and distance to transit alongside density, diversity, and design, and reported generally modest average travel elasticities. The separation of these dimensions suggests that a nearby transit stop and a concentration of homes should not be collapsed into a single measure of compactness. At the same time, modest averages do not mean that the dimensions are unimportant in every location or for every outcome. For this article, the appropriate synthesis is to regard density and proximity as related but analytically separable conditions, and to examine their observed relationship before making any inference about the broader performance of the places in which they coincide.

Debate over the magnitude of compact-development effects further cautions against selecting only literature that favors a preferred planning conclusion. Stevens (2017) interpreted his meta-regression as showing generally small effects on driving and advised against relying on compact development alone. Ewing and Cervero (2017) disputed aspects of that interpretation, emphasizing methodological choices, context, and the incompleteness of judging urban development through driving alone. This disagreement does not permit the present study to adjudicate the full compact-city debate, because it does not measure the relevant behavioral or welfare outcomes. It does establish a useful standard of interpretation: report the magnitude and scope of the observed relationship, retain competing explanations, and avoid treating either a positive association or a weak coefficient as a complete policy verdict.

Evidence that explicitly models residential density together with household travel shows why research design matters to substantive interpretation. Brownstone and Golob (2009) estimated a joint model using California travel-survey data and addressed both self-selection and missing observations related to endogenous variables. Their analysis supports a relationship between density and vehicle use within that empirical setting, but its variables, population, and outcome differ from a national database of neighborhood characteristics. The lesson for the present article is therefore methodological rather than a borrowed effect size. Density estimates acquire meaning through the process that generated the observations and the assumptions used to analyze them; an association with stop distance should be interpreted within its own spatial and measurement context.

Proximity to rail illustrates how a seemingly direct transport indicator can stand for a bundle of associated conditions. In a household study around northern New Jersey stations, Chatman (2013) found that factors including housing characteristics, density, bus service, and parking were more informative about automobile ownership and use than rail access alone. The finding does not establish that rail infrastructure is generally unnecessary, because it arises from a particular set of places and an observational

design. It does show why nearby infrastructure cannot be assumed to account for all differences between station-area and other households. Applied here, the distinction supports treating recorded transit proximity as one spatial attribute whose practical importance depends on conditions beyond the presence of the nearest stop.

Proximity service and accessibility

The choice of a transport indicator should follow the question being asked. Handy and Niemeier (1997) approached accessibility measurement as a problem of connecting academic concepts to practical policy evaluation, rather than selecting a measure solely because it was available. A nearest-stop indicator can answer a clear question about geographic separation, but it cannot answer every question about the opportunities reachable from a home. Increasing the complexity of the indicator is not automatically an improvement if it becomes less transparent or introduces assumptions unrelated to the study's objective. The defensible approach is to retain a clear proximity measure while making its conceptual boundary explicit: it describes one component of access, whose interpretation must remain narrower than the broader idea of accessibility.

Accessibility itself brings land use and transport into a common assessment. Geurs and van Wee (2004) reviewed accessibility measures against criteria including theoretical grounding, interpretability, communication, and data requirements, showing the tradeoffs involved in operational evaluation. A simple measure may be easy to compare while omitting important aspects of opportunity, whereas a more elaborate measure may better represent a construct at the cost of additional assumptions. This tension does not require abandoning readily interpretable indicators. It requires matching claims to their content. A density-proximity relationship can show that residential development and recorded stops are spatially aligned; demonstrating that residents can reach suitable employment or services would require information about destinations, travel conditions, and the relevant constraints on their journeys.

The possibility that transport access influences residential development is also longstanding. Hansen (1959) connected metropolitan residential development with accessibility and the availability of developable land, while acknowledging limitations in the operational model he presented. This early formulation is important because it reverses a simple density-first interpretation: accessibility can help shape the location of housing, rather than merely follow an existing residential concentration. It does not establish the direction or strength of feedback in the counties studied here. Instead, it supports a conceptual model in which transport and residential development may evolve together, making their observed correspondence consistent with several historical sequences. The cross-sectional evidence must therefore remain compatible with both housing-following-transit and transit-following-housing explanations.

Distance to a stop also differs from the service available at that stop. Two residences can be equally close to recorded transit while facing very different waiting times, operating hours, transfer requirements, and onward connections. These examples describe a measurement distinction rather than a claim that such differences have been quantified in the present sample. Proximity remains useful because a journey generally requires reaching a point of entry, yet the ease of that first step does not determine the performance of the full journey. The conceptual synthesis is to distinguish entry into the network from movement through it: the former can be studied with a proximity indicator, while the latter requires service and destination information that must not be presumed from distance alone.

The geographic representation of a residential area creates another gap between an indicator and lived access. Kwan (2012) identifies uncertainty about the spatial and temporal contexts that actually influence individuals, distinguishing this issue from changes caused by alternative aggregation schemes. Applied to this study, a block group is a consistent reporting unit, but its boundary is not necessarily a pedestrian catchment and its representative location is not every resident's home. This application is an inference from Kwan's methodological argument, rather than a finding about the Smart Location Database. It suggests that area-level comparisons can be useful while remaining imperfect approximations

10.5281/zenodo.24161079

of individual circumstances, and that the analysis should describe the proximity assigned to places without claiming to observe every household's actual path to transit.

Selection feedback and causal interpretation

Residential self-selection complicates claims that neighborhood characteristics determine travel behavior. Handy et al. (2005) compared cross-sectional and quasi-longitudinal evidence in Northern California and found that accounting for attitudes changed the interpretation of observed relationships, while changes over time still supported a role for the built environment. Their findings challenge both an unqualified environmental explanation and the opposite assertion that preferences necessarily explain everything. The present study does not observe attitudes or household relocations, and its outcome differs from individual travel. It therefore cannot resolve self-selection directly. The appropriate inference is that neighborhood associations describe the joint geography of people, housing, and infrastructure without isolating the preferences or decisions that produced that geography.

The requirements for stronger causal inference are more demanding than the presence of a plausible mechanism. Mokhtarian and Cao (2008) reviewed alternative approaches to residential self-selection and emphasized the importance of temporal ordering, nonspurious relationships, and designs capable of evaluating change. Statistical adjustment can help compare places that share measured characteristics, but it cannot guarantee that their unmeasured histories or institutional conditions are equivalent. This point is especially relevant when density, employment, street design, and transit may all be consequences of earlier development decisions. The synthesis is to use adjustment to clarify the observed association while explicitly distinguishing that purpose from identification of the effect of adding housing or introducing transit service.

Caution about causality should not become an automatic dismissal of observational evidence. Cao et al. (2009) reviewed 38 studies and found that built-environment associations generally remained after efforts to account for residential self-selection, although practical magnitudes were often insufficiently assessed. That body of evidence supports taking the built environment seriously while also recognizing the limitations of the designs used to study it. For the present analysis, the relevant balance is between two errors: overstating a cross-sectional relationship as a policy effect and understating a reproducible spatial pattern as meaningless merely because it is not causal. A carefully bounded comparison can establish where conditions coincide and where they diverge, even when further research is needed to explain why.

Transit provision also introduces selection on the supply side. Agencies can choose routes and stops partly in response to concentrations of potential passengers, while development decisions can respond to existing or anticipated services. A positive density-proximity association would therefore be consistent with a planning response to demand, a development response to transport, or a common response to other aspects of urban centrality. These are competing interpretations to be considered, not processes directly identified by the database. Their coexistence explains why the same empirical sign can support several historical accounts. The study's conceptual model consequently treats the association as a joint spatial outcome and reserves claims about the dominant mechanism for evidence that observes decisions or changes over time.

Comparative expectations

These arguments lead to a qualified expectation that residential density will often covary with recorded transit proximity, while the strength and form of that relationship will depend on context and measurement. The counterexpectation is equally substantive: dense areas can be poorly aligned with recorded transit, and close stops can occur where housing is sparse or where service quality remains unknown. The synthesis is an empirical research agenda centered on the distribution of alignment and mismatch, rather than on a universal density threshold. The following analysis evaluates that agenda using explicitly defined indicators and selected counties, with the aim of producing evidence that is reproducible, interpretable, and sufficiently bounded to support further planning inquiry.

Data and comparative design

Case selection and the object of comparison

The empirical design compares six metropolitan counties: New York County, New York; Cook County, Illinois; Los Angeles County, California; King County, Washington; Harris County, Texas; and Miami-Dade County, Florida. Selection precedes the estimation of density-proximity relationships and seeks variation in metropolitan setting, geographic scale, and the balance between central and peripheral residential environments. It is a purposive comparative design rather than a probability sample. Its value therefore lies in examining whether a relationship survives changes of context, not in estimating a nationally representative average. A finding repeated across these cases is evidence of comparative consistency within the selected settings; it is not a population parameter for all United States counties.

New York County provides a deliberately demanding comparison because its jurisdiction represents a concentrated central part of a much larger metropolitan system. The other selected counties encompass substantially broader mixtures of neighborhoods and municipalities. Treating these administrative units as interchangeable cities would conflate geography with urban structure, and treating each as an entire metropolitan region would omit important commuting connections. The analysis consequently uses the county as a grouping variable and the census block group as its observation. County comparisons describe the distribution of observed neighborhoods within a jurisdiction; they do not rank the performance of complete regional transit networks.

This distinction also changes the interpretation of unusually strong or weak county relationships. A jurisdiction whose neighborhoods are almost all near transit may exhibit little association between additional density and stop distance, even when its average proximity is excellent. A more geographically extensive county may display a steep gradient because its sample includes both central neighborhoods and peripheral settlements. Neither pattern, by itself, proves superior planning or a different behavioral response to density. The comparative purpose is to distinguish the level of recorded proximity from the strength of its association with residential concentration, thereby avoiding a misleading equation between a large regression coefficient and a successful transport system.

Data provenance and measurement

The study uses the 2021 Smart Location Database version 3.0 (U.S. Environmental Protection Agency, 2021), retaining the source checksum, a local six-county extract, and executable analysis code. The source is a historical release, not a measurement of conditions at the time of this analysis. Housing and population inputs principally represent 2014-2018 American Community Survey estimates; employment represents 2017; transit inputs principally target early 2020. These vintages overlap imperfectly. The release can support a comparative spatial analysis, but its components cannot establish whether housing preceded a transit improvement or whether either variable describes present-day service. Literature eligibility is separately restricted to publications issued before December 1, 2024.

D1A measures housing units per acre of unprotected land, with a flagged land-area fallback in exceptional cases. D4A measures minimum network walking distance from a population-weighted block-group centroid to a recorded transit stop, using a three-quarter-mile initial search radius. The -99999 sentinel can indicate search or data-coverage limitations. D4C summarizes scheduled service per hour over the 4-7 p.m. weekday period for routes stopping within 0.4 kilometers of the block-group boundary. The population centroids originate in 2010, whereas other inputs are newer (Chapman et al., 2021). These definitions fix what can be interpreted: the analysis relates a housing concentration measure to a representative-point measure of transit proximity, and uses service frequency only as a distinct supplementary outcome.

Housing units are the principal density numerator because the research question concerns residential development, but housing units and residents are not interchangeable. Two neighborhoods with the same number of dwellings per acre can differ in occupancy, household size, vacancy, and daily travel needs. Conversely, population weighting changes whose

neighborhood conditions receive the greatest influence without changing the physical definition of D1A. Keeping these choices separate prevents a common interpretive error: an association for housing density is not automatically an association for population density, and a population-weighted comparison is not a direct observation of individual residents' walking distances.

The denominator requires equivalent care. Unprotected acreage excludes designated protected land, but it is neither the area occupied exclusively by residential parcels nor a direct measure of legally developable land. The same residential count can yield different density values when the acreage definition changes. This study therefore treats D1A as a particular standardized indicator and evaluates a density alternative based on total land area. Agreement across denominators would support the stability of the association; disagreement would show that the meaning of density, rather than housing quantity alone, materially shapes the result.

The centroid-based outcome offers comparability at the cost of within-neighborhood detail. It assigns one origin to an entire block group, so a nearby stop for that origin does not imply a nearby stop for every address, nor does a long measured distance establish that every resident is poorly served. This limitation can be especially consequential when housing is distributed unevenly across a large block group. The resulting estimates should be read as associations between area characteristics. Address-level walking catchments, station entrances, and pedestrian barriers would be needed to translate them into precise counts of people with practical stop access.

Sample construction and missing information

The residential analytical universe retains block groups with positive population, positive housing-unit counts, positive residential density, and usable land-area information. This rule aligns the sample with inhabited residential environments and avoids taking logarithms of zero density. It also excludes some employment-dominated places that may be important to a transit network. That exclusion is appropriate for the residential question but narrows the population to which results apply. The analysis does not infer that excluded places are irrelevant; it treats workplace access as a related question requiring a different denominator and a different study population.

Geographic identifiers are reconstructed from the separately recorded state, county, tract, and block-group components. This is necessary because numeric identifiers displayed in scientific notation can lose the precision needed to distinguish neighboring areas. Correct reconstruction supports both unique-record checks and clustering at the tract level. Nonetheless, a technically valid identifier does not resolve historical boundary differences or temporal inconsistencies in the underlying variables. The reconstruction protects the integrity of this cross-sectional extract; it does not convert the release into a harmonized longitudinal database or justify comparisons with later boundaries without additional geographic reconciliation.

Unresolved transit distances are retained as information about the measurement process rather than silently removed from the entire analysis. A conventional complete-case correlation describes only neighborhoods with a usable distance, potentially understating the difference between dense and sparse environments if less dense neighborhoods disproportionately lack such records. Coding every unresolved value as a very long distance would impose an equally strong, unsupported assumption. The study instead reports the proportion with a recorded distance and then examines distance conditional on a valid record. These are complementary descriptive estimands with different denominators, not interchangeable versions of one outcome.

Recorded proximity within 400 and 800 meters supplies an additional, readily interpretable comparison. These thresholds are analytical conventions for this study, not claims about universal walking tolerances or official service adequacy. For the all-residential denominator, an unresolved observation contributes no recorded qualifying distance, while remaining explicitly unresolved in the coverage audit. This produces a transparent measure of documented proximity, not a definitive estimate of true access. The distinction permits comparisons without disguising missing information as a verified absence of service and

avoids presenting an arbitrary threshold as a behavioral law.

Analytical strategy

Descriptive and conditional comparisons

County summaries report density distributions, the fraction of block groups with usable transit-distance observations, and the distribution of distance among those observations. Medians are useful because a small number of exceptionally dense or distant areas can dominate an arithmetic mean. Medians, however, also compress distributional differences and provide no direct information about how many residents occupy the neighborhoods on either side of the middle. The paper therefore combines block-group summaries with population-weighted recorded-proximity shares and density-band comparisons. These measures answer different questions and are interpreted together rather than collapsed into a single county score.

Rank correlations describe whether denser neighborhoods tend to have shorter recorded distances without requiring a linear relationship in the original units. Their advantage is resistance to extreme magnitudes, but ranking does not remove omitted variables, selection into the valid-distance sample, or differences between counties. A pooled correlation is therefore followed by county-specific estimates. If the pooled pattern is substantially stronger, cross-county contrasts contribute to the aggregate relationship; if county-specific patterns remain strong, the relationship also characterizes within-county ordering. This comparison distinguishes scales of association without constituting an exact within-between decomposition or establishing that either association is causal.

The principal regression representation uses the base-two logarithm of residential density. A one-unit increase then corresponds to a doubling of housing units per acre, which makes coefficients easier to interpret than a one-unit change across places with sharply different starting densities. Such scaling does not make a doubling feasible everywhere or guarantee a constant response across the entire range. The coefficient summarizes an average conditional gradient over the observed data. Density-band summaries and county-specific estimates are used alongside it to identify whether that average conceals saturation, sparse support, or substantively different patterns.

The first regression family models a recorded-proximity indicator, including the indicator for a recorded stop distance of at most 800 meters. Linear probability models express the density coefficient as a percentage-point difference associated with doubling density. Their interpretation is direct, but their fitted values can fall outside the unit interval and their constant marginal slope can be inappropriate near universal coverage. Accordingly, these models are descriptive approximations, with their prediction ranges reported in the reproducibility materials. Observed proportions and county-specific patterns carry particular interpretive weight wherever the model approaches the boundaries of a probability.

Logistic sensitivity models test whether the binary associations persist when fitted probabilities are constrained to the unit interval. The recorded-distance and 800-meter models omit New York because all its eligible observations are successes, producing complete separation for its county intercept; the 400-meter model includes all six counties. These models retain the adjusted covariates and tract-clustered score-based uncertainty estimates. Their odds ratios provide a functional-form check, but the changed five-county estimation population and the distinction between odds and probabilities prevent treating them as direct replacements for six-county percentage-point coefficients.

The second family models the natural logarithm of one plus distance among records with nonnegative D4A values. Adding one retains reported zero values, which can arise from network snapping. The logarithmic transformation reduces the influence of larger distances. The coefficient describes the difference in expected log distance plus one associated with doubling density; exponentiation compares fitted geometric means of distance plus one, not arithmetic mean distances. More fundamentally, the conditional sample omits unresolved observations. The model therefore cannot describe how far all residents are from transit and cannot be treated as a substitute for the first-stage recorded-proximity comparisons.

County fixed effects remove differences in average outcome levels across the six jurisdictions and estimate the density gradient from within-county variation. This makes a pooled result less dependent on contrasting a dense central county with an extensive suburban county. Fixed effects do not, however, remove neighborhood-level historical investment, local zoning, station siting, household sorting, or route-placement decisions. The adjusted coefficient remains an association among neighborhoods within counties. Its main analytical contribution is to distinguish the geography of the comparison from the measured neighborhood relationship, rather than to supply an experimental estimate of densification.

Additional models introduce employment density, weighted intersection density excluding auto-oriented intersections, and employment-household entropy to assess how much of the residential association overlaps with other built-environment characteristics. These covariates can capture shared urban centrality and network structure, but they may also be mechanisms through which development and transit evolve together. A smaller residential-density coefficient after adjustment therefore does not show that density has no policy relevance. It shows that the association attributable to residential density alone is weaker when specified features of the surrounding environment are held constant. Both minimally adjusted and extended models are retained because they answer different descriptive questions. The adjusted specification uses log2(D1A), ln(1 + D1C), ln(1 + D3B), and untransformed D2A_EPHHM, together with an intercept and five county indicators.

The treatment of uncertainty is deliberately narrower than the language of statistical significance often implies. Census tracts define clusters for sandwich standard errors, allowing residual dependence among block groups within a tract. This does not eliminate dependence between adjacent tracts or throughout a transit corridor, and the purposively selected counties do not constitute independent random draws from a national population. The sandwich estimator uses the finite-sample multiplier G/(G - 1) × (N - 1)/(N - K), with G tracts, N observations, and K coefficients; 95% intervals use a normal critical value. Reported intervals are consequently model-based indicators of coefficient precision under the clustering convention. They should not be read as comprehensive measures of uncertainty about causality, measurement error, geographic transferability, or future policy performance.

Sensitivity and interpretive checks

Population-weighted models ask how associations change when neighborhoods containing more residents contribute more strongly to estimation. The unweighted analysis instead gives each qualifying block group equal influence. Neither is automatically superior: equal weighting treats small areas as the analytical units, while population weighting approximates the distribution of residents across those areas. The latter still assigns all residents the same block-group distance and cannot recover household-specific travel conditions. Comparison of both estimates is thus a sensitivity check on the target of description, not a method for converting aggregate observations into individual behavioral evidence. A separate sensitivity assigns each county equal total weight within the relevant model sample, using the inverse of its number of included observations.

Sensitivity checks also address the density denominator, exceptional density flags, unusually high density values, and the influence of New York County. Excluding a distinctive case is useful for testing whether it dominates a pooled pattern, but removing an extreme county merely because it weakens a preferred result would be inappropriate. The exclusion is interpreted symmetrically: stability supports a broader cross-case association, whereas change reveals dependence on sample composition. Similarly, trimming extreme observations can show whether a slope rests on a small tail, but the untrimmed results remain essential because those observations may represent real urban environments rather than data errors.

The association with service frequency is examined as a supplementary test of whether proximity and scheduled supply move together. A neighborhood can have a nearby stop with infrequent departures, and a large block group can accumulate many departures around its boundary

without providing short walks from every residence. A frequency association can therefore corroborate one aspect of density-transit alignment without validating the proximity measure as a complete index of accessibility. The purpose of the supplementary model is to expose possible divergence between indicators, not to combine them into a composite score whose weighting would conceal the underlying tradeoffs.

The analysis is exploratory and explanatory rather than preregistered or predictive. County choice and the principal measurement distinctions were established before estimating the reported relationships, but the study does not claim an independently registered specification or a held-out forecasting exercise. Multiple related estimates can reveal stability, yet they also create opportunities to overemphasize whichever specification is most favorable. To reduce that risk, the accompanying materials retain the full set of specified comparisons, sample counts, and transformations. Interpretation follows the pattern across estimates and the measurement constraints rather than a single coefficient or probability threshold.

Results

The analytical population and recorded coverage

The selected counties contribute 16,748 records from the national release of 220,740 block groups. Applying the residential eligibility rules retains 16,580 observations in 5,604 census tracts, representing 26,393,343 residents and 10,262,426 housing units in the source estimates. The 168 excluded records are not all uninhabited: together they contain 42,204 residents and 56 housing units, illustrating why positive population and positive housing counts are distinct requirements. Their institutional or other residential circumstances cannot be established from these totals alone. The results consequently describe the explicitly defined housing-based sample, rather than every person or every block group in the selected jurisdictions.

Among eligible observations, 14,156 have a nonnegative D4A value and 2,424 remain unresolved. Thus, recorded distance coverage is 85.4%, while the shares with recorded distances within 400 and 800 meters are 52.8% and 79.0%, respectively. Valid distances range from 0 to 1,207 meters. The median is 327.2 meters, but that median and the continuous-distance models describe only this bounded, selected subset. Reporting it alone would suppress the unresolved 14.6% of the analytical population of block groups. Together, the coverage share and conditional median show that relatively short recorded distances coexist with a substantial set of neighborhoods whose distance cannot be interpreted in the same way.

The distinction between unresolved distance and absent service is visible within the extract itself. Of the 2,424 eligible observations without a usable D4A value, 585 nevertheless have a nonnegative D4C service-frequency value. Conversely, 304 observations have a usable distance but no usable frequency value. These discrepancies reflect indicators with different measurement requirements and spatial footprints; they should not be interpreted as a verified count of database mistakes. Their analytical significance is direct: replacing every unresolved distance with “no transit” would contradict available evidence that the frequency indicator records service near some of those block groups. Separate status reporting is therefore necessary, not merely a precautionary choice of language.

County contrasts in levels and distributions

New York County stands apart in both residential intensity and documented proximity. Its median density is 92.84 housing units per acre, compared with 6.63 in Cook, 5.64 in Los Angeles, 4.83 in Miami-Dade, 3.56 in King, and 3.16 in Harris. Its median valid distance is also the shortest, at 176.4 meters. These contrasts establish that the central New York case occupies a different part of the observed distribution, but they do not supply a metropolitan performance ranking. Its much narrower jurisdiction and limited low-density support make it especially useful for investigating saturation, while limiting simple comparisons of its average with those of geographically broader counties.

Table 1. Residential density and recorded transit proximity by county

County	Eligible block groups	Median density HU/acre	Valid distance %	Median valid distance m	Recorded ≤800 m %
New York	1,074	92.84	100.0	176.4	100.0
Cook	3,983	6.63	89.4	283.0	84.5
Los Angeles	6,388	5.64	86.6	360.8	80.5
King	1,419	3.56	84.5	376.9	73.6
Harris	2,142	3.16	60.9	435.9	53.3
Miami-Dade	1,574	4.83	94.3	384.9	84.4

Density and valid distance are unweighted block-group medians. Recorded 800-meter proximity uses all eligible block groups; unresolved records do not count as qualifying records. Valid-distance medians exclude unresolved records. Source: original calculations from the 2021 EPA Smart Location Database.

Recorded distance coverage differs more sharply than conditional medians alone suggest. Every eligible New York block group has a valid distance, compared with 94.3% in Miami-Dade, 89.4% in Cook, 86.6% in Los Angeles, 84.5% in King, and 60.9% in Harris. Conditional median distances outside New York range from 283.0 meters in Cook to 435.9 meters in Harris. Although Harris has the longest median, its larger distributional distinction is the large unresolved share omitted from that median. The combination of measures identifies a broader difference in documented spatial reach than a comparison of typical covered neighborhoods could reveal.

The county ordering also changes with the proximity definition. Miami-Dade has a larger valid-distance share than Cook, yet Cook has a larger recorded 400-meter share, 63.7% compared with 49.0%. At 800 meters the two are close, at 84.5% and 84.4%. This is not an inconsistency: having a recorded stop within the database's search conditions and having a particularly short recorded walk are different outcomes. The comparison suggests that Miami-Dade's documented reach is relatively broad while Cook's covered distribution includes more very short distances. It would be misleading to declare either county unambiguously more accessible without specifying the indicator and the destinations or service conditions of interest.

Population weighting reveals another substantive distinction. Harris's recorded 800-meter share falls from 53.3% of eligible block groups to 42.3% of residents assigned their block group's indicator. Cook falls from 84.5% to 81.6%, Los Angeles from 80.5% to 79.0%, Miami-Dade from 84.4% to 81.4%, and King from 73.6% to 72.6%; New York remains at 100%. The Harris difference indicates that more populous block groups receive less favorable recorded proximity than an equal-block-group summary implies. It does not establish that precisely 57.7% of residents live more than 800 meters from a stop, because centroid assignment and unresolved records prevent that individual-level interpretation.

Within-county gradients and saturation

The pooled Spearman correlation between density and valid distance is -0.477. County-specific correlations are -0.518 in King, -0.430 in Cook, -0.384 in Los Angeles, -0.354 in Miami-Dade, and -0.231 in Harris, whereas New York's correlation is 0.023. The five negative associations support the expectation that denser neighborhoods often have shorter documented walks within a county. New York qualifies that expectation rather than overturning it: a jurisdiction with uniformly favorable recorded proximity need not show an additional negative density gradient. The result distinguishes variation in local ordering from the overall level of provision.

County-specific density quartiles make the distributional contrast clearer. From the lowest to highest quartile, recorded 800-meter proximity

rises from 55.9% to 99.5% in Cook, 27.6% to 82.1% in Harris, 33.8% to 98.3% in King, 57.1% to 93.4% in Los Angeles, and 66.8% to 93.9% in Miami-Dade. New York remains at 100% in every quartile. These comparisons are based on within-county ranks, so the highest quartile in one county is not an equivalent absolute density category in another. Their value is to demonstrate how strongly documented proximity differentiates neighborhoods within each local density distribution, while exposing the absence of such differentiation in the saturated New York case.

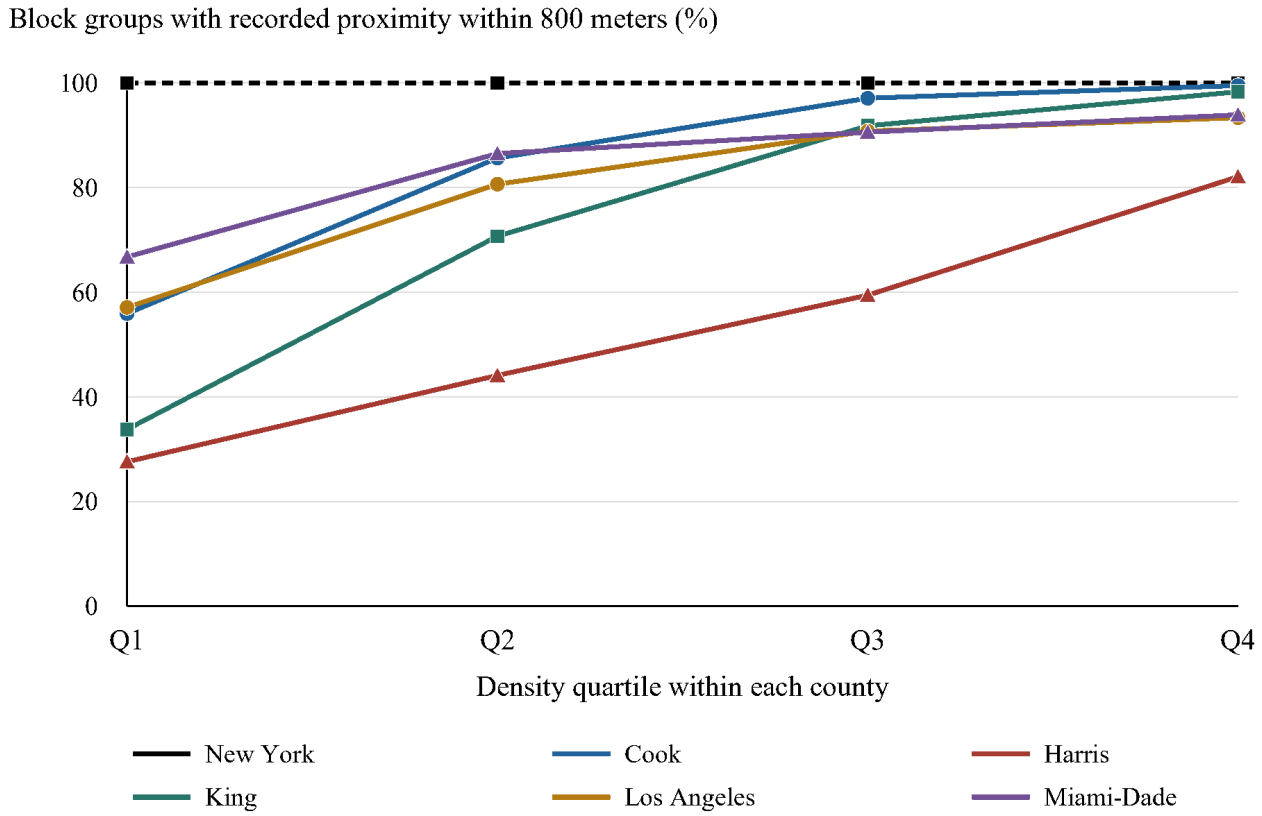


Figure 1. Recorded proximity across within-county residential-density quartiles

Quartiles are defined separately within each county and do not represent common absolute density levels. The denominator is all eligible block groups within each quartile; unresolved distances do not count as qualifying proximity. The flat New York series represents 100% in each quartile. Source: original calculations.

The quartiles also show why a common threshold cannot be read from these data. Harris's highest density quartile has a median of 7.71 housing units per acre and an 82.1% recorded 800-meter share, while King's highest quartile has a median of 9.87 and a 98.3% share. Cook's second quartile has a median of 4.85 and an 85.6% share. These are grouped descriptive comparisons with different within-group distributions and other characteristics, not matched observations or threshold estimates. Nevertheless, they demonstrate that a density level alone cannot reproduce the observed proximity pattern across jurisdictions. Local network structure and the broader urban configuration remain necessary parts of the interpretation.

Regression estimates and their scope

The pooled conditional-distance model estimates a coefficient of -0.2248 for log2 residential density. Adding county fixed effects reduces its magnitude to -0.1712, and adding employment density, weighted intersection density, and employment-household entropy yields -0.1373, with a tract-clustered 95% interval of [-0.1574, -0.1171]. This attenuation indicates that part of the pooled association overlaps with county context and the included built-environment characteristics. It does not identify a mediated share or isolate an intervention effect. The remaining coefficient establishes that the negative density-distance association persists among neighborhoods sharing the specified county and covariate conditions.

Table 2. Conditional distance models

Specification	Density coefficient	95% interval	N	R^2
Pooled	-0.2248	[-0.2414, -0.2083]	14,156	0.091
County fixed effects	-0.1712	[-0.1882, -0.1542]	14,156	0.099
County effects and controls	-0.1373	[-0.1574, -0.1171]	14,156	0.117
Adjusted population weighted	-0.1424	[-0.1677, -0.1171]	14,156	0.120

Outcome is ln(1 + D4A), restricted to nonnegative distances. Coefficients describe a doubling of D1A. Controls are ln(1 + D1C), ln(1 + D3B), and D2A_EPHHM. Intervals use tract-clustered CR1 standard errors; all models contain 4,953 tracts. R^2 is weighted for the weighted model.

Exponentiating the adjusted coefficient gives a fitted geometric mean of distance plus one that is approximately 12.8% lower for a doubling of density, with the corresponding interval approximately 11.1% to 14.6% lower. This transformation is not an estimate of the reduction in arithmetic mean walking distance after constructing additional homes. The adjusted model explains about 11.7% of variation in the logged outcome, leaving substantial neighborhood variation outside the specification. The association is therefore sufficiently consistent to characterize a spatial pattern but too limited to predict a particular site's proximity or to substitute for local network assessment.

For recorded 800-meter proximity, the adjusted six-county linear projection gives a coefficient of 0.0931, or 9.31 percentage points per density doubling, with a 95% interval of 8.75 to 9.86 points. Its fitted values range from -0.647 to 1.819, and 2,061 of 16,580 predictions fall outside the probability interval. Those departures are substantial and preclude using the model as a calibrated probability forecast. Its coefficient is retained only as a transparent average linear association. A bounded nonlinear comparison is necessary before drawing conclusions about whether the binary association depends on this restrictive functional form.

The corresponding adjusted logistic model for the five counties with variation in the 800-meter outcome yields an odds ratio of 2.04 per density doubling, with a tract-clustered 95% interval of [1.92, 2.18]. New York is excluded from that model because every eligible observation meets the threshold and its county intercept would diverge; the 400-meter model retains all six counties and yields an odds ratio of 1.94 [1.84, 2.04]. These results support a positive association under a bounded response function. The odds ratios are neither risk ratios nor percentage-point changes, and the five-county estimate does not apply to New York's already saturated outcome.

Sensitivity and supplementary service evidence

The negative conditional-distance association persists across several changes in specification. The adjusted coefficient is -0.1424 under population weighting, -0.1382 after excluding New York, -0.1377 after excluding flagged density observations, -0.1413 after trimming the highest density percentile, and -0.1402 using housing units per total land acre. Equal weighting of counties within the valid-distance sample produces a somewhat smaller coefficient of -0.1181. Stability of sign across these comparisons supports the presence of a broad spatial relationship, while the magnitude change under equal-county weighting shows that the pooled coefficient partly reflects the relative contribution of differently sized counties.

Table 3. Sensitivity of the adjusted conditional distance association

Adjusted sensitivity	Density coefficient	95% interval	N
Exclude New York	-0.1382	[-0.1579, -0.1185]	13,082
Exclude density flags	-0.1377	[-0.1578, -0.1175]	14,139
Trim highest density percentile	-0.1413	[-0.1617, -0.1210]	13,990
Housing / total land acres	-0.1402	[-0.1594, -0.1210]	14,156
Equal total county weight	-0.1181	[-0.1434, -0.0927]	14,156
Exclude zeros; ln(distance)	-0.1428	[-0.1537, -0.1319]	13,650

All models include county indicators and the three controls. The density predictor is log2(D1A), except for the total-land-density alternative. The final row models ln(distance) for positive distances; other rows model ln(1 + distance). Equal-county weights are recalculated within each model sample. Intervals are tract-clustered.

Zero distances do not explain the principal finding. The valid sample includes 506 reported zero values that may reflect network snapping; excluding them and modeling the logarithm of strictly positive distance produces an adjusted coefficient of -0.1428 [-0.1537, -0.1319]. A model in original meters gives an adjusted linear slope of approximately -59.1 meters per density doubling, although that linear model also permits implausible predictions and is not used for site forecasts. Different transformations therefore preserve the direction of association while changing its precise interpretation. The appropriate conclusion is robustness of the broad relationship, not equivalence of coefficients drawn from differently transformed outcomes.

The supplementary frequency analysis also finds a positive density association. Among 14,437 eligible observations with usable D4C values, the adjusted coefficient for log2 density in the log-one-plus-frequency model is 0.1455 [0.1290, 0.1619]. This corresponds to about 15.7% higher fitted geometric mean frequency plus one per density doubling, conditional on the included covariates. It provides additional evidence that residential concentration aligns with a dimension of scheduled supply, but D4C's aggregate spatial footprint differs from a single nearest stop. The result cannot establish headways at that stop, service reliability, or the destinations reachable by a resident; it supports examining multiple transit dimensions rather than substituting one for another.

Discussion

What the empirical pattern adds

The results support a conditional alignment interpretation of metropolitan residential development. In five counties, higher density accompanies both greater recorded proximity and shorter valid distances, and the pooled conditional relationship remains after adjusting for the specified contextual characteristics. Yet the New York case shows that a weak local slope can coexist with uniformly favorable proximity, while Harris shows that a relatively modest conditional-distance gradient can coexist with a large gradient in whether proximity is recorded at all. Density is therefore informative in different ways across the cases. The synthesis is a two-dimensional interpretation: evaluate the breadth of documented proximity and the distribution of distance within covered locations before judging the strength of housing-transit coordination.

The comparison also changes what counts as a useful counterexample. New York is not evidence that density is generally irrelevant, because its sampled neighborhoods occupy a high-density, short-distance range with little scope for further measured improvement. Conversely, Harris's unresolved observations are not evidence that all those neighborhoods lack transit, because some have independently recorded frequency information. Each case limits a different overgeneralization. The first warns against equating statistical slope with system quality; the

second warns against equating data absence with service absence. Together they show why careful interpretation can yield a more substantive conclusion than simply reporting a negative pooled correlation.

The adjusted associations locate questions for planning without specifying a single intervention. A residual negative density-distance relationship is compatible with transit following concentrated housing, housing responding to transit, and both responding to historically developed centers. The regression cannot select among these mechanisms. Its practical contribution is to identify the degree and form of alignment that any explanatory account must address. A subsequent local study could examine whether an apparent mismatch reflects route design, pedestrian barriers, constrained housing supply, or incomplete source data. That sequence uses comparative evidence to focus investigation while leaving intervention choice contingent on the mechanism actually observed.

From residential concentration to useful transit

Residential concentration can strengthen the economic case for transit because more potential passengers occupy a given service area. Guerra and Cervero (2011) found that population and employment densities were associated with both higher ridership and higher capital costs, while greater density tended to improve cost effectiveness when costs were expressed per rider or passenger mile. This finding supports coordinating land development with transit investment, but it does not supply a universal residential-density threshold. Their analysis concerns fixed-guideway investments, incorporates employment as well as population, and links guidelines to project costs. A housing-units-per-acre measure therefore cannot be substituted directly for their combined density concepts or transferred unchanged to every bus corridor. For this study, the defensible inference is narrower: the spatial alignment of homes and transit is a relevant planning condition whose consequences depend on service, employment, and costs. The observed density-distance relationship identifies where those spatial conditions tend to coincide; it cannot determine whether a particular investment is financially justified.

Distance to a stop offers an intuitively understandable measure of transit proximity, and standardized thresholds make comparisons reproducible. Yet the behavioral meaning of a threshold changes with the service reached and the people making the journey. Using Montréal travel-survey evidence, El-Geneidy et al. (2014) found that walking catchments varied across transit modes and with route, trip, household, personal, and neighborhood characteristics. Their findings challenge treating every resident inside a conventional buffer as equally served or every resident outside it as excluded. The methodological implication is to interpret threshold-based coverage as a geographic indicator, with results tested at more than one cutoff where the data permit. Even a stable density association across cutoffs would establish robustness to those particular definitions, not prove that the selected distances represent actual walking tolerance in every county. The synthesis is to retain proximity for its transparency while withholding behavioral interpretations that require local pedestrian-network and traveler evidence.

The usefulness of transit also depends on when services operate and which destinations they connect. Owen and Levinson (2015) modeled aggregate commute mode shares using accessibility to employment calculated continuously across departure times, demonstrating an approach that represents the scheduling dimension more directly than a single distance measure. Their study supports examining the opportunities reachable through the network, although an aggregate predictive model does not identify the effect of moving an individual household closer to a stop. This distinction matters when interpreting the Smart Location Database: a short access distance is only one component of a feasible journey. A nearby stop may require a long wait, an indirect route, or a transfer that defeats the purpose of the trip. Accordingly, proximity can be treated as an entry condition to the network, while employment accessibility provides a different assessment of what the network makes possible. Agreement between these measures would strengthen a descriptive interpretation;

disagreement would identify a substantive planning problem.

Operating service is a plausible mechanism through which residential concentration becomes useful accessibility. In a longitudinal analysis covering 25 North American transit agencies, Boisjoly et al. (2018) emphasized the importance of service investment, particularly bus operations, in explaining ridership patterns. The study offers evidence against assuming that urban form alone sustains transit use, but its agency-level observational relationships cannot be imported as block-group causal effects. For county comparisons, this means that dense places with nearby stops should not automatically be described as adequately served. The positive scheduled-frequency association reported here does not establish service adequacy. Favorable geography may coexist with infrequent service, limited operating hours, or insufficient capacity. Conversely, additional service along an existing corridor might improve travel opportunities without changing the measured distance to the nearest stop. Land-use and service indicators should therefore be read together where available, and any apparent lack of improvement in proximity should be interpreted in light of the indicator's limited sensitivity to operational change.

These distinctions imply that planning responses should follow the type of mismatch identified, rather than a single county ranking. A dense residential area located far from observed transit invites examination of access barriers, stop locations, and route coverage; a dense area near transit but poorly connected to destinations invites a different investigation of service and network structure. Low-density areas near useful transit may warrant attention to whether additional housing can be accommodated, whereas low-density areas with weak service raise questions about both existing residents' needs and the consequences of future development. These are diagnostic possibilities, not findings established merely by assigning observations to categories. Averages can conceal the relevant combination of conditions, and database errors can imitate apparent mismatches. The practical use of the analysis is therefore to prioritize local investigation and compare plausible interventions. It should not convert a regional statistical association into an automatic prescription for rezoning, route expansion, or service withdrawal at a particular site.

Distributional interpretation and residential stability

Concentrating public investment where many households live can increase the number of potential beneficiaries, but that objective does not exhaust the meaning of transport equity. Pereira et al. (2017) argue for evaluating accessibility as a human capability, with attention to minimum opportunities, disadvantaged groups, and the distribution of benefits and burdens. This perspective challenges the proposition that equal distances to transit necessarily constitute fair provision. People with different physical abilities, resources, or daily obligations may face different constraints even when they live equally close to the same stop. The present comparison can establish spatial patterns relevant to such an assessment, yet density and distance alone do not measure those capabilities. A defensible equity interpretation therefore asks which groups experience limited opportunities and what barriers produce that limitation. It does not label a county equitable simply because its average proximity is favorable, or inequitable solely because its settlement pattern is dispersed.

Comparative evidence also cautions against assuming that social disadvantage invariably coincides with the poorest transit geography. Foth et al. (2013) found that the most disadvantaged census tracts in their Toronto analysis had comparatively favorable job accessibility and transit travel times in both periods studied. This result demonstrates that infrastructure can serve populations with substantial social need, while their assessment also distinguishes spatial, temporal, and employment-related dimensions of transit benefits. The implication for the selected metropolitan counties is not that the same distribution must occur, but that it must be measured. An apparently favorable relationship between density and proximity can coexist with either progressive or regressive distribution across social groups. Moreover, observed service advantages do not show whether remaining needs are met. Equity analysis should consequently compare relevant population groups and

destinations within each county before drawing conclusions from the county's overall position, and any such extension should state clearly which dimensions remain unobserved.

An average relationship can also conceal a disadvantaged minority whose circumstances differ from the regional pattern. Allen and Farber (2019) found that lower-income Canadian neighborhoods generally had better transit accessibility, yet substantial numbers of low-income residents still experienced poor access. Their study illustrates why a favorable group average is not evidence that transport poverty has disappeared. Transferring its Canadian estimates to United States counties would be inappropriate, but its analytical lesson is directly relevant: the lower-access tail warrants separate examination. Population-weighted statistics may describe the experience of a typical resident more closely than an unweighted average of block groups, while neither automatically captures the most constrained population. For the present study, differences between weighted and unweighted results should be understood as changes in whose geography receives emphasis. Harris illustrates the importance of this distinction in the present results. A fuller assessment would report both broad patterns and the number or share of residents in unfavorable conditions, using verified population and disadvantage measures rather than inferring need from density alone.

Affordability creates another separation between physical proximity and effective access. El-Geneidy et al. (2016) incorporated transit fares into accessibility measures and showed that estimates based only on travel time counted more reachable employment opportunities than measures that also accounted for monetary cost. The effects differed across socioeconomic groups, indicating that financial barriers cannot simply be assumed to affect everyone equally. Nevertheless, their Montréal findings do not establish fare burdens in the counties examined here. The relevant conclusion is a limit on interpretation: a household can live near a transit stop while lacking an affordable connection to its destinations. If monetary costs are absent from the empirical specification, the analysis should retain the terminology of proximity or measured accessibility rather than claiming comprehensive access. Subsequent policy appraisal would need local fare structures, transfer rules, and household resources to determine whether a proposed improvement provides a usable opportunity to the intended beneficiaries.

Transit-oriented development introduces a further tension because improved location advantages can change housing-market demand. Padeiro et al. (2019), reviewing 35 quantitative studies, found some support for transit-related gentrification while emphasizing methodological weaknesses and the importance of local conditions and accompanying policies. Their review does not justify treating gentrification as an inevitable result of transit investment, nor does it justify dismissing residents' concerns because causal estimates are difficult. For an analysis of current density and proximity, the main implication is that favorable spatial conditions do not reveal how those conditions were produced or who can retain access to them. Residential concentration close to transit could represent long-established neighborhoods, recent construction, selective population change, or combinations of these processes. Distinguishing them requires evidence through time. The appropriate synthesis is to evaluate housing development and residential stability together when interpreting the policy relevance of proximity, while keeping the present cross-sectional findings separate from claims about neighborhood transformation.

Gentrification and displacement also require different empirical questions. Zuk et al. (2018) argue that the two have often been conflated and identify methodological limitations, including the availability and timing of data, that complicate analysis of public investment. An increase in neighborhood income does not by itself document the involuntary departure of particular residents, while stable aggregate characteristics do not establish that vulnerable households remained. The Smart Location Database's geographic measures cannot resolve either process without additional housing and household evidence. This limitation should shape the use of the paper's findings: locations that combine strong transit advantages with potential housing pressures can be prioritized for further assessment, but they should not be described as displacement cases on that basis. Preserving affordable access is a

legitimate policy objective; demonstrating the effectiveness of a particular protection requires a separate evaluative design. This distinction allows the article to take distributional risks seriously without presenting unobserved harms or remedies as empirical results.

Limitations and a research agenda

The central limitation is causal identification. Density, transit infrastructure, service allocation, employment concentration, and residential preferences can develop jointly over decades. A cross-sectional model cannot establish the sequence of those changes even when it includes several covariates and county indicators. The associations are valuable for diagnosing spatial alignment and for identifying where further investigation is warranted. They are insufficient for predicting the effect of a zoning reform, a new route, or a housing project. A causal extension would require an identifiable intervention, credible comparison areas, pre-intervention observations, and explicit consideration of anticipatory development and spillovers.

Selection into the valid-distance sample creates a second limitation that cannot be removed by reporting robust standard errors. Conditional distance regressions describe neighborhoods for which the database supplies a distance, while unresolved neighborhoods may differ systematically in density, peripheral location, or source-data coverage. The two-part presentation makes that difference visible but does not solve it statistically. A stronger extension would obtain independently audited stop inventories and pedestrian networks for all sampled counties, then calculate distances under a common unrestricted procedure. Until that is done, documented proximity and conditional distance should remain separate conclusions.

Geographic aggregation introduces uncertainty even if every underlying record is correctly measured. County boundaries partition metropolitan systems unevenly, and block groups vary in area and in the spatial distribution of housing. Re-estimating on a common grid or within comparable station catchments could change both the density numerator and the representative origin. Such changes would not necessarily invalidate the present findings; they would identify the scale at which those findings hold. A useful next study would compare block-group results with address-weighted network catchments, retaining the same transit date and distinguishing boundary effects from substantive differences in service.

The available uncertainty calculations address residual dependence only partially. Tract clustering allows nearby observations within a tract to share unobserved influences, but routes, labor markets, and land-use regimes often span multiple tracts. More elaborate spatial covariance estimators would require explicit distance or adjacency assumptions, and their precision could still obscure systematic measurement problems. The responsible interpretation therefore combines interval estimates with specification changes and the size of the observed contrasts. Future work should examine residual spatial autocorrelation and spatially robust inference before using small differences between coefficients as evidence of meaningful geographic divergence.

Temporal mismatch limits both historical interpretation and contemporary application. An older population centroid can misrepresent the location of subsequent housing development, while a transit schedule can change faster than the built environment. The assembled snapshot may consequently pair residential conditions and transit conditions that were never simultaneously observed with identical precision. This does not erase the broad spatial comparison, but it weakens fine-grained claims about particular developments or current service gaps. A longitudinal design should preserve source dates, rebuild origins from contemporary residential locations, and separate structural network change from short-term changes in service provision.

The study also cannot evaluate distributive fairness from density and stop distance alone. It lacks direct observations of fare burden, disability access, household travel schedules, route reliability, housing insecurity, and the destinations particular residents need to reach. Population weighting provides a different descriptive perspective but does not supply these missing dimensions of disadvantage. Equity conclusions should therefore be framed as priorities for additional assessment rather than findings about who benefited or suffered. Linking accessibility

calculations to relevant population characteristics and longitudinal housing outcomes would turn those priorities into questions that can be tested.

Finally, comparative evidence from six selected counties cannot determine a universal residential-density threshold for transit investment. The statistical relationship is conditioned by the observed cases, their existing networks, their geographic boundaries, and the particular density measure used. A threshold chosen from this sample could simply encode current service-allocation practices, including past underprovision. The defensible contribution is a set of measured spatial relationships and a method for interpreting them without conflating proximity, service, and accessibility. Investment decisions require additional operational, distributional, and institutional evidence that the present analysis deliberately does not manufacture.

Conclusion

Residential density and recorded transit proximity are consistently associated across much of the selected six-county sample, but the relationship is neither uniform nor self-explanatory. In 16,580 eligible block groups, the adjusted conditional-distance model associates a doubling of housing density with a 12.8% lower fitted geometric mean of distance plus one, while recorded-proximity comparisons reveal substantial differences in coverage and within-county gradients. These findings are robust to several alternative weighting, sample, and measurement choices. They remain associations within a historical, mixed-vintage database, and neither their statistical precision nor their consistency establishes what would happen if housing density were changed through policy.

The comparison's most instructive cases are the ones that qualify a simple density-first account. New York combines very high residential intensity with universal recorded 800-meter proximity and almost no within-county distance gradient. Harris combines a much lower documented-proximity share with a substantial difference between population-weighted and unweighted coverage. These patterns cannot be reduced to one universal density threshold or one county ranking. They support evaluating the existing level of proximity, the portion of the residential geography with usable evidence, and the local gradient together. The relevant planning question is how housing and transit are aligned in a particular setting, not whether one coefficient is large enough to serve as a general rule.

The broader contribution is a reproducible way to use public spatial data without asking an indicator to establish more than it measures. Density can identify concentrations of potential demand, distance can describe entry to the transit network, and service or accessibility measures can address other parts of the journey. None substitutes for evidence about affordability, distribution, residential stability, or the effects of a proposed intervention. Coordinated housing and transport planning can draw on the documented alignment while investigating those additional conditions. This synthesis gives the empirical findings practical value precisely because it identifies both the relationships that are observed and the further evidence required to explain and act on them.

References

Allen, J., & Farber, S. (2019). Sizing up transport poverty: A national scale accounting of low-income households suffering from inaccessibility in Canada, and what to do about it. Transport Policy, 74, 214-223. https://doi.org/10.1016/j.tranpol.2018.11.018

Boisjoly, G., Grisé, E., Maguire, M., Veillette, M.-P., Deboosere, R., Berrebi, E., & El-Geneidy, A. (2018). Invest in the ride: A 14 year longitudinal analysis of the determinants of public transport ridership in 25 North American cities. Transportation Research Part A: Policy and Practice, 116, 434-445. https://doi.org/10.1016/j.tra.2018.07.005

Brownstone, D., & Golob, T. F. (2009). The impact of residential density on vehicle usage and energy consumption. Journal of Urban Economics, 65(1), 91-98. https://doi.org/10.1016/j.jue.2008.09.002

Cao, X., Mokhtarian, P. L., & Handy, S. L. (2009). Examining the impacts of residential self-selection on travel behaviour: A focus on empirical findings. Transport Reviews, 29(3), 359-395. https://doi.org/10.1080/01441640802539195

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. Transportation Research Part D: Transport and

Environment, 2(3), 199-219. https://doi.org/10.1016/S1361-9209(97)00009-6

Chapman, J., Fox, E. H., Bachman, W., Frank, L. D., Thomas, J., & Rourk Reyes, A. (2021). Smart Location Database: Technical documentation and user guide (Version 3.0). U.S. Environmental Protection Agency. https://www.epa.gov/sites/default/files/2021-06/documents/epa_sld_3.0_technicaldocumentationuserguide_may2021.pdf

Chatman, D. G. (2013). Does TOD need the T? On the importance of factors other than rail access. Journal of the American Planning Association, 79(1), 17-31. https://doi.org/10.1080/01944363.2013.791008

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., & Surprenant-Legault, J. (2014). New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas. Transportation, 41(1), 193-210. https://doi.org/10.1007/s11116-013-9508-z

El-Geneidy, A., Levinson, D., Diab, E., Boisjoly, G., Verbich, D., & Loong, C. (2016). The cost of equity: Assessing transit accessibility and social disparity using total travel cost. Transportation Research Part A: Policy and Practice, 91, 302-316. https://doi.org/10.1016/j.tra.2016.07.003

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. Journal of the American Planning Association, 76(3), 265-294. https://doi.org/10.1080/01944361003766766

Ewing, R., & Cervero, R. (2017). "Does compact development make people drive less?" The answer is yes. Journal of the American Planning Association, 83(1), 19-25. https://doi.org/10.1080/01944363.2016.1245112

Foth, N., Manaugh, K., & El-Geneidy, A. M. (2013). Towards equitable transit: Examining transit accessibility and social need in Toronto, Canada, 1996-2006. Journal of Transport Geography, 29, 1-10. https://doi.org/10.1016/j.jtrangeo.2012.12.008

Geurs, K. T., & van Wee, B. (2004). Accessibility evaluation of land-use and transport strategies: Review and research directions. Journal of Transport Geography, 12(2), 127-140. https://doi.org/10.1016/j.jtrangeo.2003.10.005

Guerra, E., & Cervero, R. (2011). Cost of a ride: The effects of densities on fixed-guideway transit ridership and costs. Journal of the American Planning Association, 77(3), 267-290. https://doi.org/10.1080/01944363.2011.589767

Handy, S., Cao, X., & Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? Evidence from Northern California. Transportation Research Part D: Transport and Environment, 10(6), 427-444. https://doi.org/10.1016/j.trd.2005.05.002

Handy, S. L., & Niemeier, D. A. (1997). Measuring accessibility: An exploration of issues and alternatives. Environment and Planning A, 29(7), 1175-1194. https://doi.org/10.1068/a291175

Hansen, W. G. (1959). How accessibility shapes land use. Journal of the American Institute of Planners, 25(2), 73-76. https://doi.org/10.1080/01944365908978307

Kwan, M.-P. (2012). The uncertain geographic context problem. Annals of the Association of American Geographers, 102(5), 958-968. https://doi.org/10.1080/00045608.2012.687349

Mokhtarian, P. L., & Cao, X. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. Transportation Research Part B: Methodological, 42(3), 204-228. https://doi.org/10.1016/j.trb.2007.07.006

Owen, A., & Levinson, D. M. (2015). Modeling the commute mode share of transit using continuous accessibility to jobs. Transportation Research Part A: Policy and Practice, 74, 110-122. https://doi.org/10.1016/j.tra.2015.02.002

Padeiro, M., Louro, A., & Marques da Costa, N. (2019). Transit-oriented development and gentrification: A systematic review. Transport Reviews, 39(6), 733-754. https://doi.org/10.1080/01441647.2019.1649316

Pereira, R. H. M., Schwanen, T., & Banister, D. (2017). Distributive justice and equity in transportation. Transport Reviews, 37(2), 170-191. https://doi.org/10.1080/01441647.2016.1257660

Stevens, M. R. (2017). Does compact development make people drive less? Journal of the American Planning Association, 83(1), 7-18. https://doi.org/10.1080/01944363.2016.1240044

U.S. Environmental Protection Agency. (2021). Smart Location Database (Version 3.0) [Data set]. https://edg.epa.gov/EPADataCommons/public/OA/EPA_SmartLocationDatabase_V3_Jan_2021_Final.csv

Zuk, M., Bierbaum, A. H., Chapple, K., Gorska, K., & Loukaitou-Sideris, A. (2018). Gentrification, displacement, and the role of public investment. Journal of Planning Literature, 33(1), 31-44. https://doi.org/10.1177/0885412217716439