Int. j. adv. multidisc. res. stud. 2023; 3(6):3047-3058

Received: 10-11-2023
Accepted: 20-12-2023

International Journal of Advanced Multidisciplinary Research and Studies

ISSN: 2583-049X

Multi-Criteria Prioritization of Rural Road Rehabilitation Using Surface Condition, Population, and Market-Access Indicators

[1] Oladimeji Basit Alaka, [2] Ridwan Tiamiyu, [3] Oluwabusola Ajayi, [4] Olayemi Bolaji
[1] Rivetline Concept Limited, Lagos, Nigeria
[2] Fountain Construction Company, Nigeria
[3] Bisaj Engineering Limited, Osogbo, Nigeria
[4] Leedigital Research and Innovation, Ibadan, Nigeria

DOI: https://doi.org/10.62225/2583049X.2023.3.6.6812
Corresponding Author: **Oladimeji Basit Alaka**

Abstract
Multi-criteria prioritization ranks candidate projects by combining indicators into a score, and the standard robustness check perturbs the weights. Using 7,068 rural road links across 7,707 km in Osun and Kwara, scored on surface condition, served population, market access, health access and flood exposure, we compare that check against a perturbation of the indicator layers themselves. The comparison that carries the paper uses the one layer whose error has been independently measured. Redrawing the surface condition grade alone, through a confusion kernel reproducing an agreement coefficient of 0.041 between two readings of the same map data, gives a Kendall rank correlation of 0.546 against the base ranking. Drawing all five weights uniformly across the entire admissible simplex gives 0.669. A single indicator, redrawn at its own measured reliability, therefore disturbs the ranking more than abandoning all knowledge of the weights. Perturbing all five layers together gives 0.435 and retains 1.32 of the base top ten against 6.21 under weight perturbation. Only 77 links, 10.9 percent of the top decile and 294 km, hold a top-decile place under both perturbations; the weight check alone certifies 112 and the layer check alone 161, so each test on its own certifies links the other rejects. We conclude that where an indicator layer's reliability has been measured and is poor, weight sensitivity is the smaller of the two questions, and that where it has never been measured the honest robustness statement is that the ranking's reliability is unknown.

Keywords: Multi-Criteria Decision Analysis, Rank Stability, Uncertainty Propagation, Rural Roads, Market Access, Composite Indicators, Nigeria

1. Introduction

Nigeria's federal roads authority reported in 2020 that a majority of the country's road network was in poor or failed condition, while the rural access index literature puts the share of the rural population living within two kilometers of a road in good condition across sub-Saharan Africa at roughly a quarter (World Bank Group 2016; Iimi *et al*. 2016) [47, 24]. An agency working against those numbers, with a rehabilitation budget and more candidate roads than money, needs a ranking.

Multi-criteria decision analysis supplies one: normalize a set of indicators, weight them, combine them into a score, and rank (Hwang and Yoon 1981; Saaty 1990; Brans and Vincke 1985) [23, 40, 11]. The method is standard in rural road programs (Roberts *et al*. 2006; Iimi *et al*. 2016; Mikou *et al*. 2019) [39, 24, 32].

The standard worry about such a ranking is the weights, because they are chosen rather than measured. The standard response is a sensitivity analysis: vary the weights, observe how much the ranking moves, and report that it does not move much (Triantaphyllou and Sánchez 1997; Mukhametzyanov and Pamucar 2018; Danielson and Ekenberg 2016) [44, 33, 16]. A substantial literature refines this, from stochastic acceptability analysis (Lahdelma *et al*. 1998; Tervonen and Figueira 2007) [28, 43] to weight elicitation (Odu 2019; Riabacke *et al*. 2012; Barron and Barrett 1996) [35, 38, 4], and a parallel literature examines rank reversal under changes of method (Belton and Gear 1983; García-Cascales and Lamata 2012; Aires and Ferreira 2018) [8, 21, 1].

That apparatus treats the indicators as known and their relative importance as uncertain. In the setting studied here one

indicator is known to be very poorly measured, and it is possible to say how poorly. Grading the same road segments from two independent tag families in the same volunteered map data yields a Cohen's agreement coefficient of 0.041, which is close to no agreement at all. A second layer, flood exposure derived from Sentinel-1 radar, was measured at an F_1 of 0.356 in open floodplain and returned no detections whatever in the canopy-heavy basin.

If an input is that uncertain, then perturbing the weights while holding the layers fixed measures the smaller of two error sources and reports reassurance. This paper measures both and compares them.

It is worth being concrete about what is at stake. A state rural roads agency with a budget for perhaps eighty kilometers of road runs a prioritization, obtains a ranked list, runs a weight-sensitivity analysis, finds the ranking barely moves, and funds the top of the list. Every step of that is defensible practice. The question is what the last step rests on when the condition indicator entering the ranking has a measured agreement of 0.041.

The setting permits the comparison. The condition layer's error is measured rather than assumed, at $\kappa = 0.041$ (Cohen 1960) [15]. The indicators are the ones agencies actually use, following the rural access tradition (Roberts *et al*. 2006; World Bank Group 2016) [39, 47]. The network is large, at 7,068 links over 7,707 km built from volunteered map data (OpenStreetMap contributors 2022; Haklay 2010) [36, 22]. And aggregation can be varied separately from weighting, so method choice, weight choice and layer error can be distinguished rather than conflated (Nardo *et al*. 2008) [34].

We put weight perturbation and layer perturbation on the same prioritization and the same scale, anchoring the comparison on the single layer whose error was independently measured. We identify the set of links surviving both perturbations and show that each test alone certifies links the other rejects. And we report a network-data finding for anyone reusing remotely sensed road products: the node identifiers delivered with the satellite flood product do not produce a routable graph, and topology had to be rebuilt.

Section 2 reviews the multi-criteria and uncertainty literatures. Section 3 states the provenance of each layer and its measured error. Section 4 gives the prioritization and the perturbation experiments. Section 5 reports which dominates, and Section 6 sets out what follows and what does not.

2. Related Work

2.1 Multi-criteria prioritization and its sensitivity apparatus

The methods used here are settled. Weighted summation, TOPSIS (Hwang and Yoon 1981; Behzadian *et al*. 2012) [23, 7], the analytic hierarchy process (Saaty 1990; Chaipetch 2022) [40, 14] and the PROMETHEE outranking family (Brans and Vincke 1985; Behzadian *et al*. 2010) [11, 6] are compared regularly and generally found to agree on the extremes and disagree in the middle (Çelen 2014; Deluka-Tibljas *et al*. 2013) [13, 17]. Applied to spatial problems the combination with geographic information systems is mature (Ligmann-Zielinska and Jankowski 2014; Feizizadeh *et al*. 2014; Akpan and Morimoto 2022) [29, 20, 2].

The sensitivity apparatus is equally settled and is largely about weights. Triantaphyllou and Sánchez (1997) [44] give the classical treatment of how far a weight must move before the ranking changes; Mukhametzyanov and Pamucar (2018) [33] and Danielson and Ekenberg (2016) [16] extend it. Weight elicitation has its own literature on obtaining better numbers from experts (Odu 2019; Riabacke *et al*. 2012) [35, 38] and on surrogate schemes that avoid asking (Barron and Barrett 1996) [4].

One strand does sample weights broadly. Stochastic multi-criteria acceptability analysis inverts the question and asks which weight vectors would make a given alternative preferred, sampling the weight space rather than perturbing a point (Lahdelma *et al*. 1998; Tervonen and Figueira 2007) [28, 43]. Our uniform draw over the simplex is that construction, and we use it here as a deliberately extreme comparator rather than as a novelty: it represents the most weight uncertainty a decision problem can carry.

Two strands come nearer to the present concern. The composite indicator literature, particularly the handbook treatment (Nardo *et al*. 2008) [34] and the uncertainty analyses built on it (Saisana *et al*. 2005; Paruolo *et al*. 2013; Becker *et al*. 2017) [41, 37, 5], takes seriously that a composite's rank is a joint function of data, normalization, weighting and aggregation, and that nominal weights differ from realized influence. And the rank reversal literature (Belton and Gear 1983; Wang and Triantaphyllou 2008; García-Cascales and Lamata 2012; Aires and Ferreira 2018) [8, 45, 21, 1] establishes that rankings are fragile to changes that ought to be innocuous. What neither does is put weight-induced instability against independently measured input error on the same problem.

2.2 Rural accessibility measurement

The rural access index counts the rural population living within two kilometers of a road in good condition (Roberts *et al*. 2006) [39]. Its definitional and data problems are documented (World Bank Group 2016; Iimi *et al*. 2016; Mikou *et al*. 2019) [47, 24, 32], and the returns to rural road investment that motivate the exercise are established elsewhere (Khandker *et al*. 2009; Dercon *et al*. 2009) [27, 18]. The index's weakness is the condition half: network extent is now well observed, its surface state is not, which is why Iimi *et al*. (2016) [24] rebuild the indicator around what can be measured. Efforts to infer condition remotely, from imagery (Cadamuro *et al*. 2019) [12] or from map geometry (Jin *et al*. 2015) [25], remain exploratory. That the condition input is the weakest is already the field's own view; this paper's contribution is to price it.

2.3 The layers used here and their error estimates

Population comes from the random-forest dasymetric method (Stevens *et al*. 2015) [42], health facility locations from the register described by Blair-Freese (2019) [9], network geometry from OpenStreetMap (OpenStreetMap contributors 2022) [36], whose completeness and quality have their own literature (Haklay 2010; Barron *et al*. 2014) [22, 3], and flood detection from Sentinel-1 split-based thresholding (Martinis *et al*. 2009) [31], whose limitations under canopy are known (Makanga *et al*. 2017) [30]. Travel-time construction follows the friction-surface tradition (Weiss *et al*. 2018) [46]. Agreement between two gradings is measured with Cohen's coefficient (Cohen 1960) [15] and rank agreement with Kendall's (Kendall 1938) [26]. Speed on unsurfaced roads draws on Eaton *et al*. (1987) [19], and the regional network context on Blanford *et al*. (2012) [10].

Two of these error estimates deserve emphasis, because the

comparison below turns on them. The condition agreement of 0.041 measures a specific and limited thing: it is the cross-tabulation of two independent tag families in the same volunteered map data, a surface and tracktype reading against a smoothness reading, on the 155 ways carrying both. It is not an inter-rater study of human surveyors and no road was visited. What it measures is how far two defensible readings of the same source disagree, which is the relevant quantity for anyone building a condition layer from that source, and is not the same as the error against ground truth, which is not measured here. The flood detector's F_1 was measured against a hydrological reference rather than against field observation of impassability, for which no record exists in either area.

3. Study Area and Data Layers

3.1 Windows and the link frame

The study covers two river-basin windows: the Osun River basin in Osun State and a stretch of the Niger floodplain in Kwara State. The windows are inherited rather than chosen. They are the two the Sentinel-1 flood detector covers, and the flood exposure indicator exists nowhere else, so extending the study area would mean carrying a layer undefined over most of it.

Within the two windows, 43,740 motorable OpenStreetMap ways were split at shared vertices into 83,270 routable edges carrying 16,311 km of centerline. Edges were merged into strokes, continuing through junctions of the same road class at a deflection of at most 60 degrees (Jin *et al*. 2015) [25], and cut at 3.0 km, giving 31,779 links, of which 7,114 are rural rehabilitation candidates. Of those, 7,068 can reach both a market center and a health facility over the network and form the analysis frame: 7,707 km, median link 619 m, mean 1,090 m, 6,268 links in the Osun window and 800 in the Kwara window. Table 1 reports the frame.

Table 1: The rural link frame in the two study windows, by area and road class

Area	Class	Links	Length (km)	Median link (m)	Population served (000s)
Osun River basin	Secondary	146	378.7	2,652	148.6
Osun River basin	Tertiary	299	707.7	2,386	312.4
Osun River basin	Local	5,357	4,673.1	539	2,512.7
Osun River basin	Track	466	438.0	723	57.0
Niger floodplain	Tertiary	36	152.8	2,907	37.0
Niger floodplain	Local	461	978.1	1,267	229.8
Niger floodplain	Track	303	378.9	986	38.6
All	All	7,068	7,707.2	619	3,336.1

Note. A link is a stroke, meaning a run of road continuing through junctions of the same class with a deflection of at most 60 degrees, cut at 3.0 km. Candidate links are of class secondary, tertiary, local or track, at least 300 m long, outside the built-up clusters of Osogbo and Ilorin, and able to reach both a market center and a health facility over the network. Of 31,779 links built from 83,270 OpenStreetMap edges, 7,114 are candidates and 7,068 of those are reachable and enter the analysis. Population served is WorldPop 2020 persons whose nearest candidate link is that link, within the Rural Access Index catchment of 2000 m, so the column sums without double counting.

3.2 The routable topology had to be rebuilt

The flood detector's segment set is delivered with node identifiers that do not support routing: joining on them leaves the Kwara network in 2,237 components, the largest carrying 75 km of 4,127. Rebuilding topology by hashing way vertices from the cached extracts recovers 99.67 percent of Osun length and 56.87 percent of Kwara length into one component. The residual Kwara split is the River Niger, which has no road bridge inside the window, so the two banks are treated as separate networks because that is what they are. We record the defect here because the ranking depends on the two network travel-time indicators, and because an attribute join succeeds silently on a graph that cannot route.

3.3 The five indicators

Four layers are taken from existing sources and one is built here. Table 2 gives each indicator, its source, its direction and the status of its error estimate.

Table 2: The five indicators, their sources, and the status of each error estimate

Indicator	Symbol	Source	Status	Median	Spread	Layer coverage
Surface condition grade	C	OpenStreetMap condition tags	Measured, agreement poor	3.00	2.86–3.00	44.2%
Served population (persons)	P	WorldPop 2020 constrained	Measured, modeled surface	215	88–460	100.0%
Travel time to market (min)	M	Network routing to clusters ≥ 5,000	Modeled, threshold assumed	13.2	8.1–22.4	100.0%
Travel time to health (min)	H	Network routing to GRID3 facilities	Modeled, speeds assumed	4.8	2.7–8.7	100.0%
Flood exposure (fraction)	F	Sentinel-1 flood detector, 2022	Measured, skill low	0.000	0.000–0.263†	97.1%

Note. $n = 7{,}068$ links. All five indicators are oriented so that a larger value means a stronger case for rehabilitation. Condition is the five-point grade read off the OpenStreetMap surface, tracktype and smoothness tags; layer coverage is the share of candidate length carrying an observed tag rather than an imputed grade. Flood exposure is the largest inundated fraction the detector returned over 15 Sentinel-1 dates in 2022; layer coverage is the share of network length that detector observed. The spread column is the interquartile range, the 25th to the 75th percentile, except in the flood exposure row marked †, which runs from the 25th to the 99th percentile instead: that indicator is zero at the 25th, the 50th and the 75th percentile alike, so its interquartile range would read 0.000 to 0.000 and would say nothing. The market center threshold of 5,000 persons and the speed table for grades 3 to 5 are analyst-set, and both are perturbed in Section 4.

Surface condition is the five-point grade read from the OpenStreetMap condition tags, where grade 1 is a sealed surface in good order, grade 3 an unsealed but maintained track, and grade 5 an unmaintained earth track. Joined to the link frame by way identifier, that layer reaches 90.2 percent of edges and 77.8 percent of length. Observed tags cover 44.2 percent of candidate length and the remainder takes an imputed grade. Its measured agreement is $\kappa = 0.041$, with the caveat on what that measures given in Section 2.

Flood exposure comes from the Sentinel-1 detector, joined by way identifier and covering 97.1 percent of network length. It is nonzero on 293 of 7,068 links, every one in the Kwara window, because the detector returned zero detections in Osun. Its measured performance is $F_1 = 0.356$ in open floodplain and an almost entirely null confusion structure in the canopied basin.

Served population allocates the WorldPop constrained raster to links. Of the 4,094,813 persons in the two windows, 81.7 percent lie within 2,000 m of a candidate link, the rural access index catchment. Those 3,344,375 persons are assigned to their nearest candidate link. The population column of Table 1 sums to 3,336,128 rather than 3,344,375, the difference being the 46 candidate links that fall outside the analysis frame.

Health access is network travel time to the nearest facility in the GRID3 register, over the rebuilt topology at condition-dependent speeds.

Market access is the layer built here, and the one that had to be improvised. OpenStreetMap maps 67 marketplaces across the two states, 19 near enough to the study network to use, and none inside the Kwara window. A market indicator built on those points would be undefined across half the study area. The reported layer therefore takes a market center to be a built-up cluster of at least 5,000 persons in the WorldPop settlement frame, giving 52 centers.

3.4 Coverage, and what it means for the perturbation

Observed condition tags cover 44.2 percent of candidate length; the remaining 55.8 percent carries an imputed grade. The perturbation in Section 4 does not distinguish the two, which is conservative in one direction and not the other: imputed grades are probably less reliable than observed ones, so redrawing both at the observed agreement understates the error over the majority of the network.

The condition kernel itself rests on a thin base. The cross-tabulation is a five-by-five table built on 155 doubly-tagged ways, of which 137 fall in a single row; the top grade row is empty and the fourth holds four observations. The kernel is therefore well determined near the middle of the scale and poorly determined at its ends.

3.5 Analyst-set quantities, and what was perturbed

Six quantities are set by the analyst. Four are perturbed in the sensitivity analysis and two are not, and the distinction matters:

- *The equal weight vector.* Perturbed, across the whole simplex and within a local neighborhood.
- *The 5,000-person market threshold.* Perturbed over 2,000 to 10,000.
- *The speed multipliers for condition grades three, four and five*, for which no Nigerian measurement was located. Perturbed over assumed spans.
- *The marginal standard deviation of 0.05 defining the local weight regime.* Not perturbed. Section 4.3 states what depends on it.
- *The 3.0 km cut and 300 m floor on link length.* Not perturbed.
- *The scale of the multiplicative population error*, $\sigma = 0.21$, set so the central 90 percent of draws spans about a factor of two. Applied but not varied.

3.6 The status of each perturbation kernel

The paper's central comparison rests on kernels calibrated to a measured error. Not every kernel is, and this section states which is which.

Calibrated to a measured error. Condition, through the cross-tabulation reproducing $\kappa = 0.041$. Flood, through the detector's measured per-area confusion structure. Health facilities, dropped independently at 0.1052, the share of the eHealth Africa register recorded as unknown or not functional.

Not calibrated to a measured error. Population, whose multiplicative scale is analyst-set because WorldPop publishes no per-cell bound this analysis could apply. Travel time, whose speed table is scaled log-uniformly over a range spanning two published speed tables, which is a defensible range rather than a measurement. Market centers, whose threshold sweep spans an analyst-set range throughout.

The comparison the paper's central claim rests on is the condition-alone result, where a single measured kernel is set against the weight simplex.

4. Methods

The main perturbation arms use 2,000 draws each; the replicate-variance check uses five independent replicate runs at 400 draws; the weight regimes of Section 4.3 use 1,000 draws; and the individual-link rank trajectories of Section 5.4 use 500 draws, because they retain a full rank vector per draw.

4.1 The prioritization

Let L index the 7,068 candidate links and $k \in \{1, \ldots, 5\}$ the indicators. Write x_{ik} for the raw value of indicator k on link i. Population enters as $log(1 + p_i)$, because served population is heavily right skewed and the untransformed variable would let a handful of periurban links dominate. Raw values are winsorized at the 1st and 99th percentiles and normalized to the unit interval,

$$z_{ik} = \frac{\tilde{x}_{ik} - \min_j \tilde{x}_{jk}}{\max_j \tilde{x}_{jk} - \min_j \tilde{x}_{jk}} \tag{1}$$

With travel-time indicators reversed so every z_{ik} increases in rehabilitation priority. The composite score and the ranking are:

$$s_i(w) = \sum_{k=1}^{5} w_k z_{ik}, w \in \Delta^4 = \left\{ w : w_k \geq 0, \sum_k w_k = 1 \right\} \tag{2}$$

$$r_i(w, Z) = |\{ j \in L : s_j > s_i \}| + 1. \tag{3}$$

The base ranking takes $w = w_0 = (0.2, \ldots, 0.2)$ and the observed layers Z_0. The top decile is the 707 highest-ranked links. Equation 3 makes the paper's structure explicit: the rank is a function of two arguments, and the sensitivity literature overwhelmingly varies the first.

4.2 Aggregation as a separate axis

Before perturbing anything we vary the rule itself, holding weights and layers fixed. TOPSIS ranks on relative closeness:

$$C_i = \frac{d_i^-}{d_i^- + d_i^+}, d_i^{\pm} = \sqrt{\sum_k^{\square} \square\, w_k^2 (z_{ik} - z_k^{\pm})^2} \tag{4}$$

With z_k^+ and z_k^- the ideal and anti-ideal levels (Hwang and Yoon 1981) [23]. PROMETHEE II ranks on the net outranking flow:

$$\phi_i = \frac{1}{|L| - 1} \sum_{j \neq i}^{\square} \square \sum_k^{\square} \square\, w_k \left[P_k(z_{ik} - z_{jk}) - P_k(z_{jk} - z_{ik}) \right] \tag{5}$$

With P_k a linear preference function (Brans and Vincke 1985) [11]. A third variant replaces min-max normalization in Equation 1 with within-indicator ranks. These give a scale against which the perturbation results can be read.

4.3 Experiment one: Perturbing the weights

Two regimes answer different questions. The *simplex* regime draws $w \sim Dirichlet(1)$, uniform over Δ^4, and asks how far the ranking moves if the weights carry no information at all. The *local* regime draws $w \sim Dirichlet(\alpha w_0)$ with α set so the marginal standard deviation of each weight is about 0.05, and asks how far the ranking moves under the narrower disagreement the surrogate-weight literature works in (Barron and Barrett 1996) [4].

That 0.05 is analyst-set and is not perturbed, and any comparison against the local regime inherits it. Halving that standard deviation would roughly double the ratio reported in Section 5.2, and doubling it would roughly halve it. We therefore treat the local regime as an illustration of how the comparison behaves in a narrow weight neighborhood, and rest the paper's claim on the simplex regime, which has no free parameter.

4.4 Experiment two: Perturbing the layers

Weights are held at w_0 and each layer is redrawn. Section 3.5 lists the analyst-set quantities entering the kernels below and Section 3.6 states which of those kernels are calibrated to a measured error.

Condition is redrawn through the row-normalized cross-tabulation of the two tag families, Laplace-smoothed so a grade never observed as a first reading falls back to the column marginal. Recomputing Cohen's coefficient on that table returns 0.0409, reproducing the measured 0.041. The kernel carries the finding that the surface reading is the optimistic one, so the perturbation carries bias as well as scatter, which is what an agreement of 0.041 means. Its effect can be stated simply: at that agreement a redrawn grade is very nearly a draw from the marginal.

Flood exposure is redrawn from the detector's measured per-area confusion structure. In the Kwara window an undetected link is truly exposed with probability 0.293 and a detected link with probability 0.378, compounded over the fifteen acquisition dates under an assumption of independence across dates. In the Osun window, where the detector found nothing, an undetected link is truly exposed with probability 0.013; the Osun structure is therefore very nearly, but not exactly, null.

Population is redrawn by sampling the catchment radius from 1, 2 and 3 km, a definitional uncertainty in the rural access literature, with a multiplicative lognormal at $\sigma = 0.21$ applied on top. *Travel time* is redrawn by propagating the perturbed condition grades through the speed table and re-routing, with the whole speed table additionally scaled log-uniformly over a factor of 1.5 either way. *Health facilities* are dropped independently at 0.1052. *Market access* is redrawn by moving the threshold across 2,000 to 10,000 persons, which regenerates the center set and re-routes.

The coupling between condition and travel time is deliberate. Condition enters the composite directly and indirectly, through the speed table determining network travel time. Drawing the two independently would let a link be graded as failed and routed at sealed-road speed, and would understate the movement a condition error produces by splitting it across two partially cancelling channels. Propagating the perturbed grades through the speed table keeps them aligned, at the cost of a full shortest-path recomputation per draw.

A fourth arm draws weights and layers together.

4.5 Measures

Each draw produces a ranking compared against the base on three quantities: Kendall's τ over all 7,068 links (Kendall 1938) [26]; the top-decile overlap, the share of the base top 707 that the perturbed ranking also places in its top 707; and the number of the base top ten retained. The last is the one a decision maker feels, because a rehabilitation program funds something close to a top-ten list.

Two attributions follow. One-at-a-time redraws a single layer with the rest at their observed values, which is where the measured kernels can be isolated from the assumed ones. And the *robust core* is the set of base top-decile links holding a top-decile place in at least 90 percent of both the weight draws and the layer draws, reported alongside the sets surviving each test alone.

4.6 What replicate variance and area variance each measure

Replicate variance is the Monte Carlo noise of the perturbation, estimated from five independent replicate runs. *Area variance* is the difference between the two study windows, which is geography. The first says whether the experiment reproduces; the second is a finding. They are reported separately in Section 5.6 and are never added.

5. Results

5.1 The base ranking, and how far the rule alone moves it

The base ranking's top decile is 707 links. Its geography is lopsided: 394 of them, 55.7 percent, fall in the Kwara window, which holds only 11.3 percent of the links. That follows from the layers, because Kwara is where the flood

indicator is nonzero and where market access is worst, and it reflects the equal weights and the market threshold of the base case.

Changing the aggregation rule, with weights and layers held fixed, moves the ranking a long way. Against the weighted sum, TOPSIS gives $\tau = 0.688$ and shares 75.1 percent of the top decile; PROMETHEE II net flow gives $\tau = 0.789$; replacing min-max normalization with ranks gives $\tau = 0.690$. Table 3 reports these.

Table 3: Agreement between the three aggregation rules and between two normalizations, at equal weights

Comparison	Kendall τ	Spearman ρ	Top-decile overlap	Top-10 kept	Median shift	95th pct shift
Weighted sum vs TOPSIS	0.688	0.827	0.751	8	656	2,871
Weighted sum vs PROMETHEE II	0.789	0.921	0.805	5	399	1,867
TOPSIS vs PROMETHEE II	0.774	0.928	0.598	4	465	1,569
Min–max vs rank normalization	0.690	0.866	0.716	3	603	2,244

Note. $n = 7{,}068$ links, weights fixed at $1/5$ each. Top-decile overlap is the share of the 707 highest-priority links shared by the two rankings; top-10 kept counts how many of one ranking's ten highest-priority links appear in the other's ten. Shift columns are absolute rank differences. PROMETHEE II uses a linear preference function with indifference threshold zero and preference threshold set to each criterion's interquartile range.

The number to note is 0.688. Switching from weighted sum to TOPSIS, a choice most applications make on grounds of familiarity, disturbs this ranking about as much as drawing the weights uniformly over the entire simplex, which gives 0.669. Whatever a weight-sensitivity analysis is defending against, it is not the analyst's own choice of formula. Fig 1 places these against the perturbation arms.

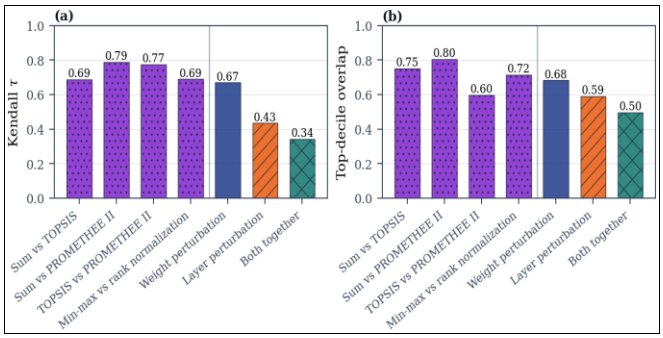

Fig 1: Aggregation rule against perturbation on a common axis, in (a) Kendall τ and (b) top-decile overlap. The four purple dotted bars compare aggregation rules on fixed weights and fixed layers: weighted sum against TOPSIS, against PROMETHEE II, TOPSIS against PROMETHEE II, and min-max against rank normalization. The three bars right of the divider are the perturbation arms. Switching from weighted sum to TOPSIS lands at $\tau = 0.69$, indistinguishable from the 0.67 produced by drawing weights across the whole simplex, so the analyst's choice of formula moves this ranking as far as abandoning the weights does. Layer perturbation at 0.43 sits below both

5.2 A single measured layer against the whole weight simplex

Table 4 reports the one-at-a-time layer attribution and Fig 2 places every source of instability on one axis.

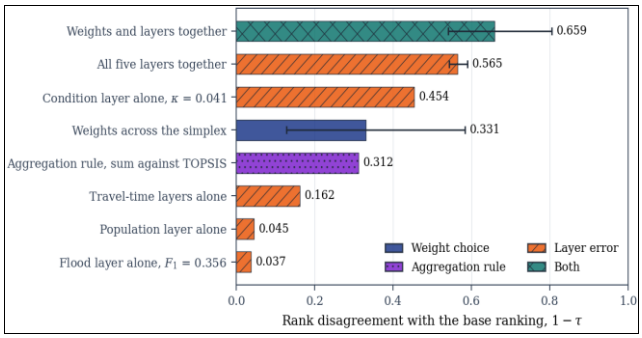

Fig 2: Every source of rank instability on one axis, as rank disagreement $1 - \tau$ against the base ranking, so longer bars mean a less stable ranking. Bars are grouped by kind: weight choice (blue, solid), layer error (orange, hatched), aggregation rule (purple, dotted) and both together (teal, cross-hatched). Error bars on the weight and joint arms span the 5th to 95th percentile of draws. The condition layer alone, at its measured $\kappa = 0.041$, reaches 0.454 against 0.331 for weights drawn across the entire simplex, which is the paper's central comparison. The flood layer is last at 0.037 because the indicator is exactly zero on 95.9 percent of links

Table 4: One layer at a time: how much of the layer-driven rank instability each indicator contributes on its own

Layer redrawn, others held fixed	Kendall τ	$1 - \tau$	SD of τ	Top-decile overlap
Surface condition, $\kappa = 0.041$	0.546	0.454	0.005	0.780
Travel time to market and to health facility	0.838	0.162	0.017	0.820
Served population	0.955	0.045	0.002	0.972
Flood exposure, $F_1 = 0.356$	0.963	0.037	0.002	0.880
All five layers together	0.435	0.565	0.014	0.591
Weights across the simplex	0.669	0.331	0.139	0.683

Note. 400 draws per layer, weights held at equal throughout. The two rows below the rule are repeated from Table 5 for comparison and use 2,000 draws. The condition layer on its own moves the ranking further than drawing the weights across the whole simplex does, which is the paper's central result. The condition perturbation is not a chosen noise level: it is the empirical cross-tabulation of two independent readings of the same 155 OpenStreetMap ways, whose exact agreement was 18.06 percent and whose κ was 0.041

Redrawing the condition layer alone, at the measured agreement, gives $\tau = 0.546$. Drawing all five weights uniformly across the entire admissible simplex gives $\tau = 0.669$. A single indicator, redrawn at its own measured reliability, moves the ranking further than abandoning all knowledge of the weights.

This is the paper's central comparison, and three properties keep it clean. Nothing about the weights is at issue: they are held at equal throughout the condition redraw. The kernel is not a chosen noise level: it is the measured cross-tabulation, and no other layer moves. And the comparator is the most extreme weight uncertainty a five-criterion problem admits, not a narrow neighborhood chosen to flatter the result.

The remaining single-layer results run travel time at $\tau = 0.838$, population at 0.955 and flood at 0.963. Travel time is partly a second channel for the same condition error, since perturbed grades propagate through the speed table, so it is not an independent effect.

The flood layer moves the ranking least, and that is not reassurance. Flood exposure is exactly zero on 95.9 percent of links, so the indicator is very nearly constant and perturbing a constant has little to move. A layer stable because it is empty is not a stable layer, and no statistic in Table 4 distinguishes the two conditions.

5.3 All five layers together

Table 5 reports the four perturbation arms and Fig 3 shows the distributions.

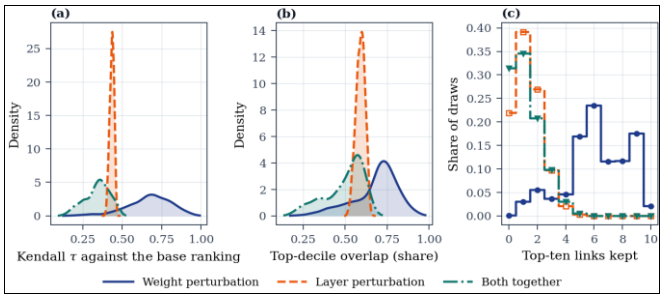

Fig 3: Distributions of the three stability measures across 2,000 draws per arm: (a) Kendall τ against the base ranking, (b) top-decile overlap, and (c) the number of the base top ten retained, plotted as the share of draws attaining each count. Weight perturbation (blue, solid), layer perturbation (orange, dashed) and both together (teal, dash-dotted). The layer distribution is not merely lower than the weight distribution but far narrower, because its kernels are fixed while the weight simplex is sampled uniformly. Panel (c) shows the practical consequence: weight draws are centered near six of ten retained while layer draws concentrate on zero to two

Table 5: Rank stability against the base ranking under weight perturbation, layer perturbation and both

	Kendall τ				Top-decile	Top-10
Source of instability	**Mean**	**SD**	**5th pct**	**95th pct**	**overlap**	**kept**
Weights drawn near equal weights	0.917	0.039	–	–	0.912	8.87
Weights drawn across the simplex	0.669	0.139	0.417	0.872	0.683	6.21
Layers redrawn within measured error	0.435	0.014	0.410	0.456	0.591	1.32
Both together	0.341	0.079	0.194	0.459	0.497	1.20
Layer loss over weight loss	1.71×				1.29×	2.29×
Draws in which layer error is worse	92.8%				75.4%	94.6%

Note. 2,000 draws of each kind, $n = 7{,}068$ links. The headline weight draws are uniform over the five-dimensional simplex, so they cover every weight vector an analyst could defend; the first row instead draws from a Dirichlet concentrated on equal weights, which is the narrower regime the surrogate-weight literature works in, over 1,000 draws. Layer draws hold the weights at equal and redraw the indicators inside their own error: condition through the cross-tabulation measured between two independent readings of the same OpenStreetMap ways, Cohen's $\kappa = 0.041$; flood exposure through the detector's measured skill, $F_1 = 0.356$ in the Niger floodplain and zero detections in the Osun basin; population through the catchment radius and a multiplicative lognormal; travel time through the speed table's own published span. The ratio row divides one minus the layer mean by one minus the weight mean, and for the top-10 column divides links lost rather than links kept.

Weight perturbation across the simplex gives mean $\tau = 0.669$, top-decile overlap 0.683, and 6.21 of the base top ten retained. Perturbing all five layers together, with weights fixed at equal, gives $\tau = 0.435$, overlap 0.591, and 1.32 retained. On rank disagreement that is 1.71 times the weight arm, on top-decile loss 1.29 times, and on loss from the top ten 2.29 times. The layer draw is the less stable of the two in 92.8 percent of paired draws.

Because every layer moves at once, that arm is an upper bound on layer-driven instability.

Against the local weight regime the arithmetic gap is much larger, at 6.8 times the rank disagreement. That figure depends on the marginal standard deviation of 0.05 defining the local weight regime, discussed in Section 4.3.

5.4 What happens to an individual link

Aggregate correlations understate what a decision maker experiences, because a program acts on specific roads. Fig 4 traces the base top-ranked links through both perturbations over 500 draws, and Fig 5 shows the decile transitions.

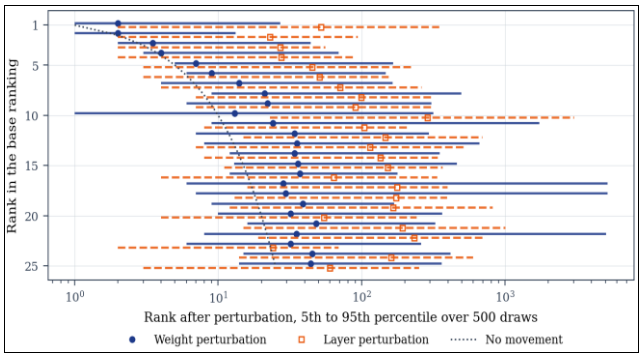

Fig 4: Rank intervals for the 25 highest-ranked links under each perturbation, over the same 500 draws. Each horizontal bar spans the 5th to 95th percentile of the perturbed rank on a logarithmic axis, with the median marked: weight perturbation (blue, solid, circles) and layer perturbation (orange, dashed, squares). The dotted line marks no movement. Layer intervals are wider than weight intervals across most of the top 25, and several links whose weight interval stays within the top ten have layer intervals reaching past rank 100

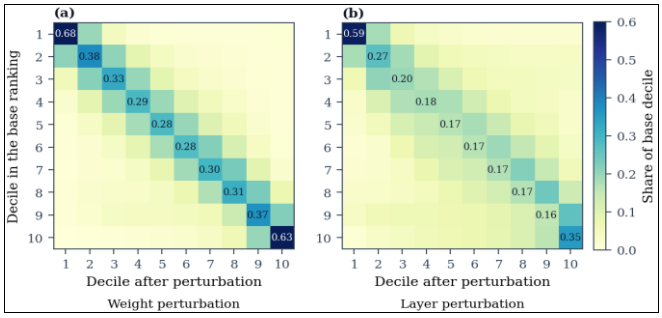

Fig 5: Decile transition matrices: the share of each base decile landing in each perturbed decile, under (a) weight perturbation and (b) layer perturbation. Under weight perturbation the diagonal holds, at 0.68 for the top decile and 0.63 for the bottom. Under layer perturbation the top decile retains 0.59 and every interior decile falls to between 0.16 and 0.27, so the middle of the ranking is close to reshuffled

The link ranked first in the base list holds first place in 2.4 percent of layer draws. The comparable weight figure is not quoted here, because the weight winner frequency was computed on a separate 1,000-draw regime run and the two are not like for like. What the rank trajectories do support, being computed on the same 500 draws for both arms, is the shape in Fig 4: the top-ranked links carry visibly wider layer intervals than weight intervals across most of the top 25, and several links whose weight interval stays inside the top ten have layer intervals reaching past rank 100.

Fig 6 shows the same thing as a probability of holding a top-decile place.

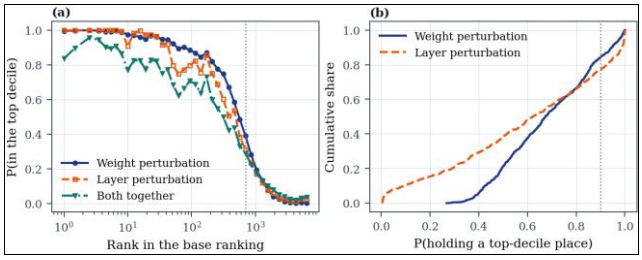

Fig 6: Top-decile acceptability. Panel (a) gives the probability that a link holds a top-decile place against its rank in the base ranking, on a logarithmic rank axis, for weight perturbation (blue, solid, circles), layer perturbation (orange, dashed, squares) and both together (teal, dash-dotted, triangles); the dotted vertical line marks the 707-link decile boundary. Panel (b) gives the cumulative distribution of that probability across base top-decile links, with the dotted line at the 0.90 threshold used to define the robust core. The layer curve in (b) lies well to the left of the weight curve, which is why the two tests certify different sets

5.5 The robust core, and what each test alone would certify

Table 6 reports the links surviving both perturbations at the 0.90 threshold: 77 links, 10.9 percent of the base top decile, 294 km of road. Relaxing the threshold to 0.80 gives 156 links and to 0.50 gives 395. Fig 7 maps them.

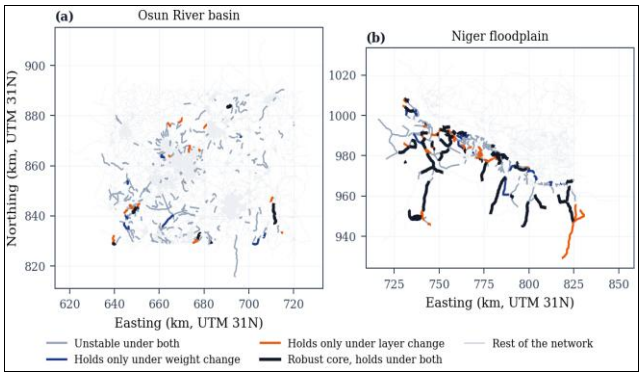

Fig 7: The robust core across (a) the Osun River basin and (b) the Niger floodplain windows, in UTM zone 31N. Base top-decile links unstable under both perturbations (slate), links holding only under weight change (dark blue), holding only under layer change (orange), and the robust core holding under both (black), against the rest of the network (light grey). 69 of the 77 core links fall in the Kwara window, inverting a frame that is 89 percent Osun by count

Table 6: The robust core, and replicate variance and area variance reported separately

	Retention threshold		
Links of the base top decile that	0.90	0.80	0.50
Holds under weight change only	112	237	557
Holds under layer change only	161	241	437
Robust core, holds under both	77	156	395
Robust core, share of top decile	10.9%	22.1%	55.9%
Robust core, length (km)	294	488	1,031
Robust core in the Osun basin	8	20	110
Robust core in the Niger floodplain	69	136	285
Replicate variance: SD of the mean τ across 5 independent replicate runs, 400 draws each			
Weight perturbation	0.0046		
Layer perturbation	0.0005		
Area variance: mean τ computed within each area			
Weight perturbation, Osun basin / Niger floodplain	0.666 / 0.653		
Layer perturbation, Osun basin / Niger floodplain	0.389 / 0.480		

Note. The base top decile holds 707 links. A link enters the robust core if it holds a top-decile place in at least the stated share of both the 2,000 weight draws and the 2,000 layer draws. Replicate variance and area variance are reported separately and must not be added: the first is the spread of the perturbation draws and is small, the second is geography and is not. The between-area gap in mean τ is 0.014 under weight perturbation and 0.091 under layer perturbation, the latter being 175 times the layer replicate standard deviation.

The operational point is that each test on its own certifies links the other rejects. At the 0.90 threshold, 112 top-decile links survive the weight test alone and 161 survive the layer test alone, while only 77 survive both. An agency running the conventional weight-sensitivity analysis and stopping there would hold 112 links it believed robust, of which 35 are not. An agency running only the layer test would hold 161, of which 84 fail the weight test. Neither test subsumes the other, and the intersection is smaller than either.

The stronger claim, that the weight test is the more permissive of the two, does not hold. At the 0.80 threshold the layer test certifies 241 links against the weight test's 237, so which test is more permissive depends on where the threshold is set. What does hold at every threshold computed is that both sets exceed the intersection, which is the reason to report the intersection.

The 0.90 threshold is a choice and the core's size is sensitive to it: 77 links at 0.90, 156 at 0.80, 395 at 0.50. That sensitivity restates the result rather than undermining it. If the ranking were stable the count would be flat in the threshold, because links would hold their place nearly always or nearly never. Instead it more than quintuples between 0.90 and 0.50, so a large middle band of the top decile holds its position somewhere between half and nine tenths of the time. An agency choosing a threshold is choosing how much of that band to fund, and the honest presentation is the curve rather than a point on it.

The core's geography inverts the network's: 69 of the 77 core links are in the Kwara window and 8 in Osun, against a frame that is 89 percent Osun by count. What survives both perturbations is concentrated where the flood layer carries any signal at all.

5.6 Replicate variance and area variance

Across five independent replicate runs at 400 draws each, the standard deviation of the mean Kendall τ is 0.0046 for weight perturbation and 0.0005 for layer perturbation. The experiment reproduces to the third decimal place, and none of the differences above is within reach of Monte Carlo noise.

Area variance is a different quantity and a larger one. Within each window, weight perturbation gives mean $\tau = 0.666$ in Osun and 0.653 in Kwara, a gap of 0.014. Layer perturbation gives 0.389 and 0.480, a gap of 0.091. On top-decile overlap the layer arm gives 0.473 in Osun against 0.353 in Kwara.

The two are reported apart and never added, because they answer different questions. Replicate variance says the experiment reproduces; it does. Area variance says the windows behave differently under layer perturbation, and the direction is instructive: Kwara is the more stable on Kendall τ and the less stable on top-decile overlap. That combination is what happens when one window's flood indicator carries real signal and the other's is very nearly constant. Pooling them into a single error bar would hide the only interesting thing either says.

5.7 Benchmarks

Nine quantities were compared against published measurement. Six of them meet a published range directly, and those six divide evenly: two fall inside the range, two sit at its edge, and two fall outside it. The remaining three admit no verdict of that kind.

Inside the range are the rural population within 2 km of a road, at 81.7 percent, and the share of length carrying an observed condition tag, at 44.2 percent. At the edge are the sealed-surface share and the population-weighted travel time to the nearest health facility, at 10.7 minutes including the walk to the link. Outside are the population within one hour of a market center, at 97.8 percent, which is above published values and is not comparable without stating that our market centers are population clusters rather than mapped marketplaces, and the frequency with which the top-ranked alternative survives a change of weights, at 46.8 percent when the weights are drawn across the simplex, which is below the published band in the direction the paper would predict.

Of the three that admit no such verdict, one is consistent in direction rather than in level: Mikou *et al*. (2019) [32] report that changing the definition of a rural access indicator moves it from 41.5 to 85.9 percent on the same country, a definitional span far wider than any weight choice, which is the same phenomenon this paper measures on a ranking rather than a headline share. One is inside by construction, the Osun flood null. And one has no available comparator, and it is the paper's own: we located no published study reporting how often a top-ranked alternative survives a perturbation of the input layers. That absence is the gap the paper is written into.

Two checks failed on first running and both were acted on. The travel-time comparison was rebuilt like for like, with the walk to the link included, because the published figures include it and ours did not. And the local weight regime was added because the published winner-frequency band was measured under near-consensus weights, so comparing it against simplex-wide draws was not a comparison.

6. Discussion

6.1 What the comparison establishes

On this network, a single indicator redrawn at its own measured agreement of 0.041 disturbs the ranking more than drawing all five weights uniformly across the entire admissible simplex, by 0.546 against 0.669 in Kendall τ. The conventional robustness check, run on this problem in the narrow regime such checks normally use, would have reported $\tau = 0.917$ and 8.87 of the top ten retained.

The claim is a narrow one. The weight literature is right about weights. What does not follow is that a robustness analysis varying only the argument that is cheap to vary, and reporting the result as robustness without qualification, has established robustness. Equation 3 has two arguments.

6.2 What generalizes and what does not

The number does not generalize. It depends on how badly the layer happens to be measured, and the condition layer here sits near the floor. A prioritization built on surveyed condition data would show a different balance and might well show the opposite one.

The design generalizes, with one caveat about its cost. Any prioritization whose layers have a measured error can run this comparison cheaply once that measurement exists. Obtaining the measurement is the expensive part, and in this case it was not obtained by fieldwork: two tag families already present in the same volunteered data were compared against each other, which costs nothing but measures disagreement between readings rather than error against ground truth.

Where no layer carries an error estimate, this comparison cannot be run, and the defensible statement is that the ranking's reliability is unknown rather than that it is robust. That is an uncomfortable thing to write in a report to a funder, which may be part of why the weight-sensitivity convention persists.

6.3 Relation to the composite indicator literature

The composite indicator literature has been close to this point without quite arriving at it. Saisana *et al*. (2005) [41] argue that uncertainty and sensitivity analysis should run jointly over data, normalization, weighting and aggregation rather than over weights alone; Becker *et al*. (2017) [5] show nominal weights and realized influence diverge sharply in practice; and Paruolo *et al*. (2013) [37] note that what a weight does depends on the variance of the variable it multiplies. All three imply input error should matter at least as much as weight choice. What has been available less often is a case where the input error is independently measured, so the two can be put on one scale. This setting supplies that: the condition agreement and the detector's performance were both measured on the source data for other purposes, so neither is our own estimate of our own layer's reliability.

The result also sits consistently with the rank reversal literature (Belton and Gear 1983; García-Cascales and Lamata 2012) [8, 21] and adds a quantity to it. Switching aggregation from weighted sum to TOPSIS moves this ranking to $\tau = 0.688$, close to a uniform draw over the entire weight simplex at 0.669. Two of the three choices an analyst makes freely, the rule and the weights, move the ranking by about the same amount, and one measured input moves it further than either.

6.4 Implications for practice

Report the intersection rather than the ranking. The 77 links surviving both perturbations are a defensible program; the 707-link top decile is not, and the 112-link weight-robust set is worse than either because it carries unearned confidence.

Treat a stable indicator as suspect until its stability is explained. The flood layer was the most stable of the five

and the least informative, being exactly zero on 95.9 percent of links, and no statistic in Table 4 separates those two conditions.

And when reusing a published spatial product, verify that it supports the operation being asked of it. The flood product's node identifiers are correct for the per-segment question they were built for and produce a shattered graph for routing, and nothing in the delivered data announces the difference.

6.5 Limitations

The condition perturbation assumes the confusion kernel has the same structure as the true error. A measured K constrains the diagonal and not the full matrix, so a systematically biased grader would produce different rank movement at the same K, and the agreement itself is between two readings of the same volunteered source rather than against ground truth, which nobody has measured for this network.

The market layer is the weakest of the five, because population clusters substitute for marketplaces rather than measuring them, and the 67 mapped marketplaces across two states are too few and too unevenly distributed to validate against.

7. Conclusions

We built a multi-criteria prioritization of 7,068 rural road links across 7,707 km in Osun and Kwara from condition, population, market access, health access and flood exposure, and perturbed it in the weights, as the literature does, and in the indicator layers.

The comparison that carries the paper uses the one layer whose error was independently measured. Redrawing surface condition alone, at a measured agreement of 0.041, gives a Kendall correlation of 0.546 against the base ranking, while drawing all five weights uniformly across the entire admissible simplex gives 0.669. One indicator at its own measured reliability outweighs complete ignorance of the weights.

Perturbing all five layers together gives 0.435 and retains 1.32 of the base top ten against 6.21 under weight perturbation.

Only 77 links, 10.9 percent of the top decile and 294 km, hold a top-decile place under both perturbations. The weight check alone certifies 112 and the layer check alone 161, so each test passes links the other rejects and the intersection is smaller than either. That intersection is what an agency should fund.

The practical recommendation is to stop reporting weight sensitivity, on its own, as robustness where an input layer's reliability is poor or unknown, and to report instead the set surviving perturbation of both arguments. Where layer error has never been measured, that fact is the robustness result, and it is more useful to a decision maker than a rank correlation of 0.917.

8. References

1. Aires Renan Felinto De Farias, Luciano Ferreira. The Rank Reversal Problem in Multi-Criteria Decision Making: A Literature Review. Pesquisa Operacional. 2018; 38(2):331-362. Doi: https://doi.org/10.1590/0101-7438.2018.038.02.0331
2. Akpan Uduak, Risako Morimoto. An Application of Multi-Attribute Utility Theory (MAUT) to the Prioritization of Rural Roads to Improve Rural Accessibility in Nigeria. Socio-Economic Planning Sciences. 2022; 82:101256. Doi: https://doi.org/10.1016/j.seps.2022.101256
3. Barron Christopher, Pascal Neis, Alexander Zipf. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Transactions in GIS. 2014; 18(6):877-895. Doi: https://doi.org/10.1111/tgis.12073
4. Barron F Hutton, Bruce E Barrett. Decision Quality Using Ranked Attribute Weights. Management Science. 1996; 42(11):1515-1523. Doi: https://doi.org/10.1287/mnsc.42.11.1515
5. Becker William, Michaela Saisana, Paolo Paruolo, Ine Vandecasteele. Weights and Importance in Composite Indicators: Closing the Gap. Ecological Indicators. 2017; 80:12-22. Doi: https://doi.org/10.1016/j.ecolind.2017.03.056
6. Behzadian Majid RB, Kazemzadeh A Albadvi, Aghdasi M. PROMETHEE: A Comprehensive Literature Review on Methodologies and Applications. European Journal of Operational Research. 2010; 200(1):198-215. Doi: https://doi.org/10.1016/j.ejor.2009.01.021
7. Behzadian Majid S. Khanmohammadi Otaghsara, Morteza Yazdani, Joshua Ignatius. A State-of the-Art Survey of TOPSIS Applications. Expert Systems with Applications. 2012; 39(17):13051-13069. Doi: https://doi.org/10.1016/j.eswa.2012.05.056
8. Belton Valerie, Tony Gear. On a Short-Coming of Saaty's Method of Analytic Hierarchies. Omega. 1983; 11(3):228-230. Doi: https://doi.org/10.1016/0305-0483(83)90047-6
9. Blair-Freese IO. Geo-Referenced Infrastructure and Demographic Data for Development. 2019 IEEE Global Humanitarian Technology Conference (GHTC), 2019. Doi: https://doi.org/10.1109/GHTC46095.2019.9033027
10. Blanford Justine I, Supriya Kumar, Wei Luo, Alan M MacEachren. It's a Long, Long Walk: Accessibility to Hospitals, Maternity and Integrated Health Centers in Niger. International Journal of Health Geographics. 2012; 11:24. Doi: https://doi.org/10.1186/1476-072X-11-24
11. Brans JP, Vincke PH. Note-a Preference Ranking Organisation Method. Management Science. 1985; 31(6):647-656. Doi: https://doi.org/10.1287/mnsc.31.6.647
12. Cadamuro Gabriel, Aggrey Muhebwa, Jay Taneja. Street Smarts: Measuring Intercity Road Quality Using Deep Learning on Satellite Imagery. Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '19), 2019, 145-154. Doi: https://doi.org/10.1145/3314344.3332493
13. Çelen Aydın. Comparative Analysis of Normalization Procedures in TOPSIS Method: With an Application to Turkish Deposit Banking Market. Informatica. 2014; 25(2):185-208. Doi: https://doi.org/10.15388/informatica.2014.10
14. Chaipetch Pawarotorn. Analytical of Multi-Criteria Approach for Identifying the Weight and Factor of Rural Road Maintenance Prioritization. International Journal of GEOMATE. 2022; 22(91):70-79. Doi: https://doi.org/10.21660/2022.91.gxi334
15. Cohen Jacob. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement.

1960; 20(1):37-46. Doi: https://doi.org/10.1177/001316446002000104
16. Danielson Mats, Love Ekenberg. A Robustness Study of State-of-the-Art Surrogate Weights for MCDM. Group Decision and Negotiation. 2016; 26(4):677-691. Doi: https://doi.org/10.1007/s10726-016-9494-6
17. Deluka-Tibljas Aleksandra, Barbara Karleusa, Nevena Dragicevic. Review of Multicriteria-Analysis Methods Application in Decision Making about Transport Infrastructure. Journal of the Croatian Association of Civil Engineers. 2013; 65(7):619-631. Doi: https://doi.org/10.14256/jce.850.2013
18. Dercon Stefan, Daniel O Gilligan, John Hoddinott, Tassew Woldehanna. The Impact of Agricultural Extension and Roads on Poverty and Consumption Growth in Fifteen Ethiopian Villages. American Journal of Agricultural Economics. 2009; 91(4):1007-1021. Doi: https://doi.org/10.1111/j.1467-8276.2009.01325.x
19. Eaton RA, Gerard S, Dattilo RS. A Method for Rating Unsurfaced Roads. Transportation Research Record (Washington, DC). 1987; 1106:34-38. https://onlinepubs.trb.org/Onlinepubs/trr/1987/1106v2/1106v2-004.pdf
20. Feizizadeh Bakhtiar, Piotr Jankowski, Thomas Blaschke. A GIS Based Spatially-Explicit Sensitivity and Uncertainty Analysis Approach for Multi-Criteria Decision Analysis. Computers & Geosciences. 2014; 64:81-95. Doi: https://doi.org/10.1016/j.cageo.2013.11.009
21. García-Cascales M Socorro, Teresa Lamata M. On Rank Reversal and TOPSIS Method. Mathematical and Computer Modelling. 2012; 56(5-6):123-132. Doi: https://doi.org/10.1016/j.mcm.2011.12.022
22. Haklay Mordechai. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design. 2010; 37(4):682-703. Doi: https://doi.org/10.1068/b35097
23. Hwang Ching-Lai, Kwangsun Yoon. Multiple Attribute Decision Making: Methods and Applications. Vol. 186. Lecture Notes in Economics and Mathematical Systems. Springer Berlin Heidelberg, 1981. Doi: https://doi.org/10.1007/978-3-642-48318-9
24. Iimi Atsushi, Farhad Ahmed, Edward Charles Anderson, *et al*. New Rural Access Index: Main Determinants and Correlation to Poverty. In Policy Research Working Paper 7876. World Bank, 2016. Doi: https://doi.org/10.1596/1813-9450-7876
25. Jin Weihua, Ziyong Zhou, Huan Wang, Mengya Wang. Urban Road Network Generalization Based on Strokes. 2015 23rd International Conference on Geoinformatics, 2015, 1-5. Doi: https://doi.org/10.1109/geoinformatics.2015.7378678
26. Kendall MG. A New Measure of Rank Correlation. Biometrika. 1938; 30(1-2):81-93. Doi: https://doi.org/10.1093/biomet/30.1-2.81
27. Khandker Shahidur R, Zaid Bakht, Gayatri B Koolwal. The Poverty Impact of Rural Roads: Evidence from Bangladesh. Economic Development and Cultural Change. 2009; 57(4):685-722. Doi: https://doi.org/10.1086/598765
28. Lahdelma Risto, Joonas Hokkanen, Pekka Salminen. SMAA - Stochastic Multiobjective Acceptability Analysis. European Journal of Operational Research. 1998; 106(1):137-143. Doi: https://doi.org/10.1016/s0377-2217(97)00163-x
29. Ligmann-Zielinska Arika, Piotr Jankowski. Spatially-Explicit Integrated Uncertainty and Sensitivity Analysis of Criteria Weights in Multicriteria Land Suitability Evaluation. Environmental Modelling & Software. 2014; 57:235-247. Doi: https://doi.org/10.1016/j.envsoft.2014.03.007
30. Makanga Prestige Tatenda, Nadine Schuurman, Charfudin Sacoor, *et al*. Seasonal Variation in Geographical Access to Maternal Health Services in Regions of Southern Mozambique. International Journal of Health Geographics. 2017; 16:1. Doi: https://doi.org/10.1186/s12942-016-0074-4
31. Martinis S, Twele A, Voigt S. Towards Operational Near Real-Time Flood Detection Using a Split-Based Automatic Thresholding Procedure on High Resolution TerraSAR-X Data. Natural Hazards and Earth System Sciences. 2009; 9(2):303-314. Doi: https://doi.org/10.5194/nhess-9-303-2009
32. Mikou Mehdi, Julie Rozenberg, Elco Koks, Charles Fox, Tatiana Peralta Quiros. Assessing Rural Accessibility and Rural Roads Investment Needs Using Open Source Data. In Policy Research Working Paper 8746. World Bank, 2019. Doi: https://doi.org/10.1596/1813-9450-8746
33. Mukhametzyanov Irik, Dragan Pamucar. A Sensitivity Analysis in MCDM Problems: A Statistical Approach. Decision Making: Applications in Management and Engineering. 2018; 1(2):51-80. Doi: https://doi.org/10.31181/dmame1802050m
34. Nardo Michela, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, Anders Hoffmann, Enrico Giovannini. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing, Paris, 2008. Doi: https://doi.org/10.1787/9789264043466-en
35. Odu GO. Weighting Methods for Multi-Criteria Decision Making Technique. Journal of Applied Sciences and Environmental Management. 2019; 23(8):1449-1457. Doi: https://doi.org/10.4314/jasem.v23i8.7
36. OpenStreetMap contributors. OpenStreetMap, 2022. https://www.openstreetmap.org
37. Paruolo Paolo, Michaela Saisana, Andrea Saltelli. Ratings and Rankings: Voodoo or Science? Journal of the Royal Statistical Society Series A: Statistics in Society. 2013; 176(3):609-634. Doi: https://doi.org/10.1111/j.1467-985X.2012.01059.x
38. Riabacke Mona, Mats Danielson, Love Ekenberg. State-of-the-Art Prescriptive Criteria Weight Elicitation. Advances in Decision Sciences, 2012, 1-24. Doi: https://doi.org/10.1155/2012/276584
39. Roberts Peter, Shyam KC, Cordula Rastogi. Rural Access Index: A Key Development Indicator. In Transport Papers TP-10. World Bank, 2006. Doi: https://doi.org/10.1596/17414
40. Saaty Thomas L. How to Make a Decision: The Analytic Hierarchy Process. European Journal of Operational Research. 1990; 48(1):9-26. Doi: https://doi.org/10.1016/0377-2217(90)90057-i
41. Saisana M, Saltelli A, Tarantola S. Uncertainty and Sensitivity Analysis Techniques as Tools for the

Quality Assessment of Composite Indicators. Journal of the Royal Statistical Society Series A: Statistics in Society. 2005; 168(2):307-323. Doi: https://doi.org/10.1111/j.1467-985X.2005.00350.x
42. Stevens Forrest R, Andrea E Gaughan, Catherine Linard, Andrew J Tatem. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS One. 2015; 10(2):e0107042. Doi: https://doi.org/10.1371/journal.pone.0107042
43. Tervonen Tommi, José Rui Figueira. A Survey on Stochastic Multicriteria Acceptability Analysis Methods. Journal of Multi-Criteria Decision Analysis. 2007; 15(1-2):1-14. Doi: https://doi.org/10.1002/mcda.407
44. Triantaphyllou Evangelos, Alfonso Sánchez. A Sensitivity Analysis Approach for Some Deterministic Multi-criteria Decision-making Methods. Decision Sciences. 1997; 28(1):151-194. Doi: https://doi.org/10.1111/j.1540-5915.1997.tb01306.x
45. Wang Xiaoting, Evangelos Triantaphyllou. Ranking Irregularities When Evaluating Alternatives by Using Some ELECTRE Methods. Omega. 2008; 36(1):45-63. Doi: https://doi.org/10.1016/j.omega.2005.12.003
46. Weiss DJ, Nelson A, Gibson HS, *et al*. A Global Map of Travel Time to Cities to Assess Inequalities in Accessibility in 2015. Nature. 2018; 553(7688):333-336. Doi: https://doi.org/10.1038/nature25181
47. World Bank Group. Measuring Rural Access: Using New Technologies. World Bank, 2016. Doi: https://doi.org/10.1596/25187