AI-MCDM Investment Scoring for Smart Cities under Carbon, Energy, and Geoeconomic Risk

El Ghali Sany

MSc Procurement & Supply Chain Management

INSEEC Business School, Lyon

May 2026

ABSTRACT

This study developed and tested an AI-enabled multi-criteria decision-analysis scorecard for ranking smart city infrastructure projects under energy-price, carbon-cost, and geoeconomic-risk uncertainty. The framework combined XGBoost-based demand forecasting, carbon-adjusted Decoupled Net Present Value, Real Options valuation, Fuzzy-AHP/MEREC weighting, and AROMAN ranking. The study then benchmarked the constructed project sample against World Bank PPI, EIB project, EIBIS, and World Bank WDI/ESG datasets. XGBoost reduced mean absolute error by 45.8 percent relative to Ridge regression in the original temporal validation. The supplementary WDI/ESG validation showed that XGBoost slightly outperformed Ridge on the external country-year panel. Carbon pricing produced no ranking change at €30/tCO_2, limited but detectable ranking effects at €85/tCO_2, and stronger disruption under the €150/tCO_2 Net Zero scenario. Geoeconomic risk was not supported at the full panel level, although Egypt produced exploratory country-specific evidence consistent with a negative GPR-investment-score relationship. The AROMAN scorecard produced investment rankings that diverged from NPV-only rankings across all 25 projects, supporting the auditability, traceability, and sensitivity disclosure of multi-criteria infrastructure investment rankings. The study concluded that AI-MCDA frameworks could support transparent infrastructure prioritization when their outputs were interpreted as decision-support evidence rather than objective optimization results.

Keywords: smart city, AI-MCDM, AROMAN, XGBoost, carbon pricing, geoeconomic risk, EU ETS, decoupled net present value, TOPSIS

1. INTRODUCTION

1.1 Background and Context

Global urbanization makes smart and sustainable city infrastructure a major capital-allocation priority. These urban transitions require large upfront investments and extended payback periods to modernize transportation, energy, and water systems (Gatti, 2024). Decision-makers must allocate constrained public and private capital while navigating systemic vulnerabilities, including volatile energy markets, decarbonization mandates, and shifting geopolitical landscapes (Magwedere & Marozva, 2023). Infrastructure valuation has historically relied on traditional financial models, especially the Discounted Cash Flow (DCF) method. Espinoza (2022) demonstrated that traditional DCF analyses only partially captured the deep uncertainty associated with long-term climate adaptation and macroeconomic shocks. By discounting future liabilities and emphasizing short-term returns, conventional DCF models can exacerbate the tragedy of the horizon, undervalue resilient investments, and support suboptimal capital allocation in long-term urban infrastructure.

1.2 Research Gap

Scholars have increasingly used Multi-Criteria Decision-Making (MCDM) methods to evaluate smart city sustainability, but existing frameworks largely treat macro-environmental risks in isolation (Alizadeh, 2017). Few quantitative models translate macroeconomic and geopolitical shocks into asset-level financial metrics. Goloshchapova et al. (2023) showed that quantitative energy-investment models often did not fully account for geopolitical uncertainty. Jin et al. (2023) argued that geopolitical risk, climate risk, and energy-market dynamics should be examined together. Artificial Intelligence (AI) and Machine Learning (ML) models improve infrastructure cost forecasting, but these predictive architectures often omit external exogenous variables such as energy-price volatility, carbon costs, and geoeconomic risk (Shamim et al., 2025). A defined research gap therefore remains: no unified theoretical or mathematical framework currently integrates AI-driven forecasting, dynamic carbon costs, energy-price scenarios, and geoeconomic risk proxies into a single investment scorecard.

1.3 Research Objectives and Hypotheses

This research designs and validates an AI-enabled, multi-criteria investment scorecard for ranking smart city infrastructure projects under simultaneous energy, carbon, and geoeconomic uncertainties. The study investigates five sub-questions: (1) How can ML algorithms enhance the accuracy of infrastructure cost and demand forecasting compared with traditional parametric models? (2) How do dynamic carbon pricing mechanisms and volatile energy scenarios alter asset-level cash flows and bankability? (3) How can news-based geoeconomic risk proxies be quantified and integrated into financial valuation models? (4) How can a hybrid MCDM framework synthesize risk-adjusted financial metrics with qualitative socio-environmental indicators? (5) To what extent does this unified scorecard improve model auditability and ranking robustness compared with single-indicator deterministic models?

H1: XGBoost models outperform linear baselines by minimizing mean absolute error and root mean square error in infrastructure demand forecasting.

H2: Dynamic carbon pricing and energy transition scenarios affect project rankings compared to zero-carbon baseline evaluations.

H3: Integration of the GPR index reduces risk-adjusted financial scores of exposed infrastructure projects.

H4: Multi-criteria scorecards improve the auditability, traceability, and sensitivity disclosure of infrastructure investment rankings compared with single-indicator NPV models.

2. LITERATURE REVIEW

2.1 Multi-Criteria Investment Evaluation of Infrastructure

Infrastructure investment evaluation requires the balancing of constrained public budgets with competing technical, economic, and socio-environmental objectives. Saaty's Analytic Hierarchy Process (AHP) structured such problems by decomposing them into hierarchies and deriving criteria weights through pairwise expert comparisons. As urban planning objectives became more multifaceted, the limitations of traditional AHP prompted hybrid models. Pujadas et al. (2017) applied the MIVES methodology, which integrated MCDM with Multi-Attribute Utility Theory (MAUT). Marcelo et al. (2015) proposed the World Bank Infrastructure Prioritization Framework (IPF) as a visual decision-support tool. Sarmas et al. (2022) developed a TOPSIS-based decision-support system that achieved equivalent energy savings while reducing investment costs by 15 percent. Abdelkader et al. (2023) designed a two-tier model that combined Fuzzy ANP with an ensemble of seven algorithms merged through the Copeland algorithm. Despite this evolution, existing multi-criteria frameworks still rely on static performance snapshots and deterministic inputs. They therefore do not fully integrate compounding macroeconomic vulnerabilities over multi-decade project lifecycles.

2.2 Carbon and Energy Risk in Project Finance

Climate-sensitive project finance requires valuation methods that capture volatile market conditions. In, Weyant, and Manav (2021) proposed a forward-looking asset-level cash-flow simulation framework that evaluated DSCR and IRR at each pay period under physical and transition scenarios. Their results showed that transition risks generally generated larger financial impacts than physical risks. Assab (2024) extended the conventional CAPM by incorporating physical climate loss through a flood-risk premium applied to the loan interest rate. This extension showed that upfront adaptation investment reduced climate exposure and lowered the cost of debt for resilient infrastructure. Richstein and Neuhoff (2022) evaluated financing costs of low-carbon investments through CVaR Monte Carlo simulation and showed that Carbon Contracts-for-Difference reduced investment risk and carbon mitigation costs. Flora and Tankov (2022) developed a Real Options approach with Bayesian learning to identify the optimal timing of green-project entry and carbon-intensive asset decommissioning. This methodological evolution supports a shared conclusion: deterministic approaches ignore deep uncertainty in energy markets and discount future climate liabilities.

2.3 Geoeconomic Risk Proxies

Geopolitical instability affects macroeconomic fluctuations and infrastructure investment behavior. Caldara and Iacoviello (2022) developed the Geopolitical Risk (GPR) index and demonstrated that heightened GPR reduced industrial production, employment, international trade, and stock returns. The index separated geopolitical threats from geopolitical acts, revealing an asymmetry between anticipated and realized shocks. Nasouri (2025) extended this framework through the GARCH-MIDAS model and demonstrated cross-country heterogeneity. Emerging markets exhibited pronounced GPR-driven financial stress, while advanced economies showed a safe-haven effect. Su et al. (2023) investigated GPR effects on global supply chains through wavelet-based Quantile-on-Quantile Regression and found that extreme climate events exerted more enduring effects than geopolitical disruptions. The literature therefore demonstrates that isolated assessment of geoeconomic risk is insufficient for robust infrastructure investment valuation. Jin et al. (2023) argued that future research should examine integrating geopolitical risks, climate risks, and energy-market dynamics; this study responds to that call.

2.4 Artificial Intelligence in Investment Scoring

Accurate infrastructure cost forecasting has historically relied on parametric estimation and traditional bottom-up approaches. Conventional techniques only partially capture nonlinear interactions among technical, organizational, and external cost drivers. Shamim et al. (2025) reported that Artificial Neural Networks were the most widely adopted AI architecture and achieved 85-99 percent accuracy. Narbaev et al. (2023) demonstrated that XGBoost achieved MAPE values of 6.22-9.70 percent, outperforming traditional index-based EVM models that generated errors of 14.35-17.43 percent. Sing, Ma, and Gu (2025) developed an enhanced cost-contingency prediction framework that integrated a complexity index into ANNs and improved R^2 from 0.808 to 0.889. Two barriers still restrict the operationalization of AI in infrastructure investment scoring. First, black-box interpretability limits auditability. Second, automated cost-estimation architectures often omit external macroeconomic variables, including volatile energy prices, carbon costs, and geoeconomic risks. This paper addresses both barriers.

3. METHODOLOGY

3.1 Research Design and Methodological Rationale

This study adopted a quantitative, post-positivist, model-building methodology to test whether an AI-enabled multi-criteria decision-analysis scorecard could rank smart-city infrastructure projects under energy, carbon, and geoeconomic uncertainty. The methodology refers to the overall research logic: a deductive design, secondary-data architecture, scenario-based valuation layer, and validation protocol. The methods refer to the specific analytical tools used within that logic: XGBoost forecasting, DNPV and Real Options valuation, Fuzzy-AHP/MEREC weighting, AROMAN ranking, Wilcoxon testing, Spearman rank analysis, OLS panel regression, GARCH-MIDAS modeling, cross-method MCDM comparison, and external benchmarking.

A quantitative model-building approach was selected because the research question required comparable project rankings, measurable forecast errors, and reproducible sensitivity tests. Qualitative interviews or case-study methods could have explained stakeholder perceptions, but they would not have tested rank sensitivity under controlled carbon-price and geoeconomic scenarios. A single DCF or NPV model was also rejected because it could not represent non-financial public-value criteria, carbon-transition exposure, or ranking auditability. The selected design therefore matched the study objective: to evaluate whether a structured AI-MCDA architecture produced more traceable investment rankings than a financial-only baseline.

3.2 Data Collection, Instruments, and Source Validity

The study used unobtrusive secondary data rather than newly generated survey or experimental data. This choice reduced respondent bias and allowed the analysis to use institutional datasets with documented provenance. Five primary data groups supported the scorecard: EMBER EU ETS carbon-price data for carbon-cost scenarios, the Caldara and Iacoviello GPR index for geoeconomic-risk proxies, World Council on City Data indicators for smart-city benchmarks, World Bank WDI/WGI indicators for macroeconomic and institutional controls, and EIB project records for infrastructure financial baselines.

The measurement instruments were public data exports, documented indices, and reproducible software tools. Python was used for data cleaning, harmonization, model fitting, ranking calculations, robustness checks, and figure generation. Reliability was addressed by preserving raw files, processed files, scripts, and validation outputs in the Zenodo replication package. Construct validity was addressed by mapping each criterion to a documented source, defining the expected direction of each criterion before scoring, and separating predictive variables from ranking criteria. External benchmarking used the World Bank PPI full STATA file, EIB financed-project exports, EIB pipeline records, EIBIS indicators, and the EIB Municipalities Survey only as plausibility checks; these files did not recalibrate AROMAN weights or replace the 25-project scoring matrix.

3.3 Population, Sampling Frame, and Inclusion Criteria

The target population was composed of smart-city infrastructure projects that require long-lived capital allocation under carbon, energy, digital, and geopolitical uncertainty. The sampling frame combined EIB project archives and the World Bank PPI database, because these sources documented infrastructure-finance projects across energy, transport, water, ICT, and urban infrastructure sectors. The analytical project sample contained 25 representative projects distributed across five domains: smart grids, EV charging networks, public transport electrification, smart water management systems, and urban data centers. Five projects were allocated to each domain across 10 countries: France, Germany, Spain, Portugal, Morocco, the Netherlands, Italy, Poland, Turkiye, and Egypt.

Projects were included when they met four criteria: they belonged to a smart-city-relevant infrastructure domain, had sufficient project-type information to support financial and non-financial scoring, could be linked to country-level macroeconomic and geoeconomic indicators, and could be evaluated under carbon-price and energy-transition scenarios. Projects were excluded when they lacked a usable infrastructure category, could not be mapped to the scoring criteria, had no plausible financial benchmark, or fell outside the study scope of smart-city infrastructure. Each project was evaluated across 13 criteria covering financial performance, environmental transition, social and smart-city contribution, and geoeconomic exposure.

The sample size was determined by matrix balance and model purpose rather than by random population inference. The 25-project matrix provided five cases per infrastructure domain, which allowed controlled comparison across project types while keeping the AROMAN ranking matrix auditable. The 190 country-year observations used for the original forecasting task and the 3,547 usable WDI/ESG observations used for supplementary validation provided the statistical basis for prediction testing. Because the project sample was purposive, the study did not claim population-level inference from the 25 projects. Statistical power remained limited for H3, where low GPR variation across European economies constrained panel-level detection; this limitation was disclosed and treated as a boundary condition rather than hidden as a design strength.

3.4 Variables and Analytical Controls

The dependent variables were defined separately for each hypothesis. H1 used electric power consumption per capita as the demand-forecasting outcome. H2 used AROMAN rank change under carbon-price scenarios. H3 used the risk-adjusted investment score as the dependent variable in the panel specification. H4 used auditability, traceability, sensitivity disclosure, cross-method concordance, and divergence from NPV-only ranking as evaluation outcomes.

The independent variables also differed by hypothesis. H1 compared model class, with XGBoost evaluated against linear baselines. H2 varied carbon-price scenarios at EUR 0, EUR 30, EUR 85, and EUR 150 per tonne of CO_2. H3 used the GPR country index as the key independent variable. H4 varied scorecard structure, criterion weights, MCDM method, and the AROMAN lambda trade-off parameter. Control variables included GDP growth, inflation, renewable energy consumption, CO_2 emissions per capita, country fixed effects, and selected WDI/WGI indicators. This variable structure separated prediction, valuation, ranking, and robustness tests, reducing the risk of conflating causal interpretation with predictive performance.

3.5 Four-Layer Evaluation Framework and Execution Procedure

A unified four-layer framework was developed to evaluate and prioritize smart-city infrastructure projects under deep uncertainty. The execution pipeline followed this sequence: data extraction -> harmonization -> demand forecasting -> carbon and financial valuation ->

criteria weighting -> AROMAN ranking -> sensitivity testing -> external benchmarking -> replication-package export. This workflow made the procedure reproducible and separated each analytical operation from the next.

Layer 1 collected and quantified macro-environmental risk inputs. Geoeconomic risk was captured through the Caldara and Iacoviello GPR index, while transition risk and carbon liability were modeled through EU ETS carbon-price trajectories. Layer 2 translated those inputs into asset-level financial metrics. DNPV separated the time value of money from project-specific climate risk, extended CAPM adjusted the cost of capital, and Real Options valuation with CVaR Monte Carlo simulation generated probabilistic expected NPV. Layer 3 structured and weighted financial and non-financial criteria using a hybrid Fuzzy-AHP and MEREC weighting scheme within a MIVES/IPF decision-tree logic. Layer 4 produced final rankings using AROMAN.

The final priority score for each project was calculated as:

$$R_i = L_i\lambda + M_i(1-\lambda), \text{ where } \lambda = 0.5$$

The two-step normalization procedure specified by Đalić et al. (2021) was applied. Linear normalization and vector normalization were computed separately for each criterion and then averaged before AHP-MEREC weights were applied. This procedure was used because single-normalization implementations produced L_i and M_i values dominated by absolute project scale and generated near-inverted cross-method concordance. The corrected procedure yielded strong cross-method concordance with TOPSIS and COPRAS. The lambda baseline of 0.5 assigned equal weight to cost minimization and benefit maximization, while sensitivity analysis across lambda values from 0.3 to 0.7 disclosed the normative dependence of final rankings.

3.6 Hypothesis Testing and Data Analysis Strategy

H1 was tested using paired temporal validation errors from XGBoost, Ridge regression, and Lasso regression. Eight walk-forward temporal validation runs were used to reduce data leakage. The Wilcoxon Signed-Rank Test compared paired MAE and RMSE distributions, with $p < 0.05$ used as the confirmation threshold. H2 was tested by comparing baseline and carbon-stress AROMAN rankings with Spearman rank correlation and Average Absolute Difference in Ranking. H3 was tested using OLS panel regression and a GARCH-MIDAS country-specific specification, with GDP growth, inflation, renewable energy consumption, CO_2 per capita, and country fixed effects used as controls. H4 was tested through AHP consistency ratios, one-at-a-time weight sensitivity, direct AROMAN versus NPV-only comparison, and cross-method concordance with TOPSIS and COPRAS.

The analytical strategy therefore matched each hypothesis to a specific statistical or ranking test. Prediction hypotheses used forecast-error metrics. Ranking hypotheses used rank-correlation and rank-difference metrics. Geoeconomic-risk hypotheses used panel and volatility specifications. Auditability hypotheses used consistency, sensitivity, traceability, and cross-method robustness evidence. Results were interpreted conservatively: predictive feature importance was treated as evidence of model contribution, not causal effect; external benchmarking was treated as plausibility evidence, not full project-level external validation.

3.7 Bias Control, Ethics, and Reproducibility

Bias-control protocols were embedded in the research design. Temporal validation prevented future observations from entering training data. Leave-one-country-out validation assessed geographic transferability. Criterion directions, carbon scenarios, and confirmation thresholds were specified before interpretation. Cross-method comparison with TOPSIS and COPRAS tested whether rankings depended on one MCDM algorithm. Weight sensitivity analysis tested whether ranking outcomes depended on single criteria or on the AROMAN lambda parameter. Randomization and double-blind procedures were not applicable because the study used secondary non-reactive data and deterministic computational procedures rather than human-subject experimentation.

Ethical risk was low because the study did not involve human participants, animal subjects, intervention, clinical data, private records, or personally identifiable information. All datasets were public or institutionally published. The main ethical obligation was therefore transparent reporting, accurate source attribution, and reproducibility. The replication package deposited on Zenodo contains cleaned analytical datasets, scripts, validation outputs, supplementary materials, and figure-generation files. It is available at https://doi.org/10.5281/zenodo.20044642.

A contingency protocol was used to address foreseeable methodological risks. Missing data were handled through documented filtering and harmonization rather than undocumented imputation. Incomplete project-level external validation was addressed through PPI/EIB plausibility benchmarking and explicit limitation wording. Low GPR variance in European economies was addressed by reporting Egypt-specific findings as exploratory rather than generalizable. Model opacity was addressed through feature-importance reporting, external validation, cross-method MCDM checks, and a public replication package.

4. RESULTS

4.1 Hypothesis 1: Machine Learning versus Linear Baseline Forecasting

To test H1, XGBoost and Ridge and Lasso linear regression models were trained on macro-environmental and geoeconomic features drawn from the master dataset (190 observations, 10 countries, 2003–2021) to forecast electric power consumption per capita as a proxy for smart-city infrastructure demand. Eight temporal walk-forward validation splits were constructed to prevent data leakage, with training sets progressively expanding from 2003–2013 through to 2003–2020. The Wilcoxon Signed-Rank Test was then applied to the paired error distributions across all runs.

XGBoost achieved a mean MAE of 479.7 kWh/capita (±190.9) across eight temporal runs, representing a 45.8 percent reduction compared with the Ridge linear baseline (MAE = 884.3 ±150.1) (Table 1; Fig. 1). The Wilcoxon test yielded $W = 36.00$ and $p = 0.0039$, confirming at the 99 percent confidence level that the observed error reduction was statistically significant and not attributable to chance. These results were consistent with Narbaev et al. (2023), who reported XGBoost performance gains in project cost forecasting.

Feature importance analysis on the full-sample XGBoost model showed that CO_2 emissions per capita (47.0%), country encoding (20.4%), and GPR country index (16.1%) were the three dominant predictors (Fig. 2). This result indicated that geoeconomic and carbon variables contributed to predictive performance within the fitted model, but it should not be interpreted as causal evidence. H1 was therefore confirmed.

Table 1. Temporal Validation Performance of XGBoost and Ridge Regression for Electric Power Consumption Forecasting.

Model	**MAE (kWh/capita, mean ± SD)**	**RMSE (kWh/capita, mean ± SD)**	**MAPE (%)**	**Wilcoxon p-value**
XGBoost	**479.7 ± 190.9**	**704.5 ± 254.3**	**13.80 ± 5.27**	**0.0039*** **
Ridge Regression (baseline)	884.3 ± 150.1	1087.3 ± 163.7	26.48 ± 7.61	—
Lasso Regression	870.3 ± 155.8	1077.9 ± 159.0	25.54 ± 7.56	0.0039***

Note. MAE and RMSE are reported in kWh per capita. SD = standard deviation. MAPE = mean absolute percentage error. The Wilcoxon p-value compares XGBoost and Ridge across temporal validation runs. Source: Author calculations.

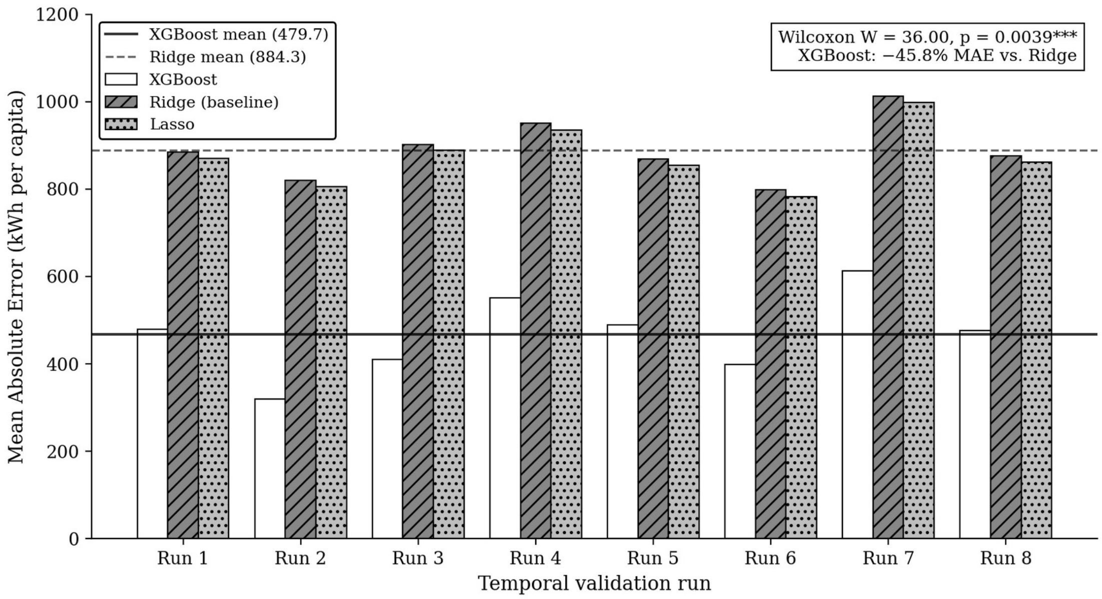

Figure 1. Model error across temporal validation runs in the infrastructure-demand forecasting task. Bars show mean absolute error (MAE, kWh per capita) for XGBoost, Ridge regression, and Lasso regression across eight walk-forward validation runs (n = 8). The horizontal reference line marks the Ridge baseline mean. The inset reports the paired Wilcoxon test comparing XGBoost and Ridge (W = 36.00, p = 0.0039), showing lower prediction error under XGBoost.

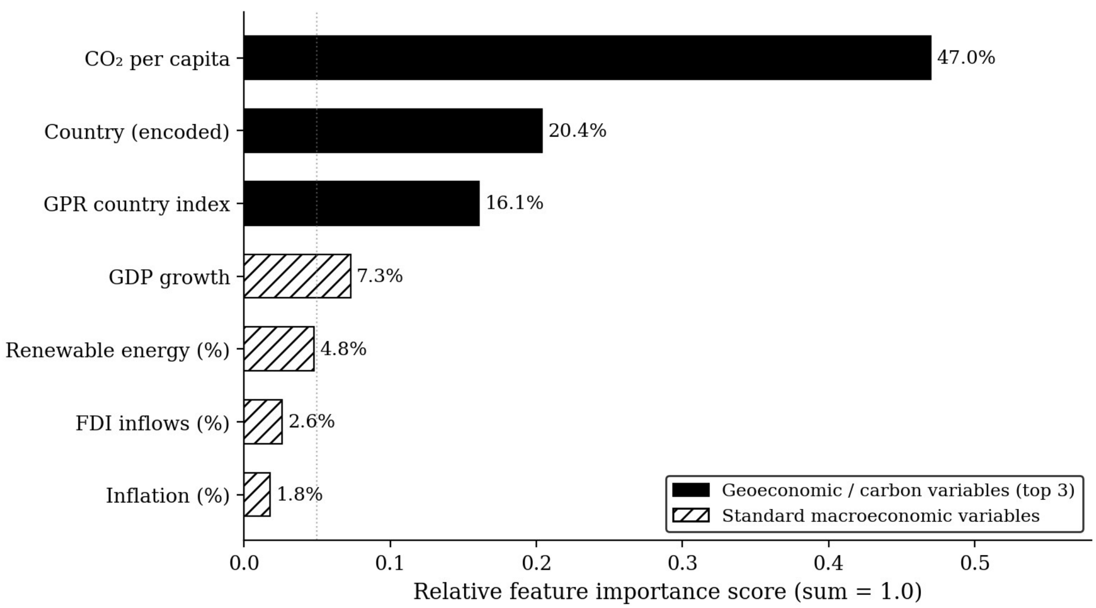

Figure 2. Feature importance from the fitted XGBoost demand-forecasting model. Horizontal bars show relative feature-importance shares normalized to sum to 1.0 for six predictors in the full-sample model. Solid bars identify geoeconomic and carbon predictors, and hatched bars identify macroeconomic controls. CO2 emissions per capita, country encoding, and the GPR country index formed the three largest contributors to predictive performance.

4.2 Hypothesis 2: Carbon and Energy Scenarios Change Project Rankings

To test H2, the AROMAN scoring formula ($Ri = Li\lambda + Mi(1-\lambda)$, $\lambda = 0.5$) was applied to the 25-project sample under four carbon price scenarios: Baseline (€0/tCO_2), Moderate (€30/tCO_2), Below 2°C (€85/tCO_2), and Net Zero (€150/tCO_2). The NPV of each project was adjusted downward by the discounted present value of carbon liabilities, calculated as the product of the carbon price, the project's ETS exposure coefficient, and a 20-year project life discounted at 8 percent. Spearman Rank Correlation Coefficients (SCC) and Average Absolute Differences in Ranking (AADR) were then computed between the baseline and each stress scenario.

At the project type level, Data Center projects demonstrated relative resilience under the Below 2°C scenario (mean rank improvement of −0.6 positions) due to their higher baseline NPV margins, while Water Management and Transport projects suffered the greatest proportional score erosion. The most penalized individual project was Water Management — Porto (P19), which fell two positions as its carbon exposure-to-NPV ratio was least favorable in the sample. These findings align with Flora and Tankov (2023), who demonstrate that transition scenario uncertainty drives differential asset valuation depending on carbon intensity. H2 was supported, with limited ranking effects at €85/tCO_2 and stronger effects under the €150/tCO_2 Net Zero scenario.

The Moderate scenario (€30/tCO_2) produced no ranking change ($\rho = 1.000$, AADR = 0.00), consistent with Richstein and Neuhoff (2022), who argued that low carbon price signals do not materially alter investment decisions when financial margins are sufficient (Table 2; Figs. 3, 4, and 5). The Below 2°C scenario (€85/tCO_2) yielded $\rho = 0.9969$, AADR = 0.24, and 5 of 25 projects re-ranked. Under the Net Zero scenario (€150/tCO_2), 12 of 25 projects were re-ranked (ρ

= 0.9838, AADR = 0.80), demonstrating a progressive, monotonic relationship between carbon price intensity and ranking disruption.

Table 2. Carbon Price Scenario Effects on Smart-City Project Rankings.

Scenario	Carbon price (€/tCO_2)	Spearman ρ	p-value	AADR	Re-ranked projects (n)
Baseline	€0	1.0000	—	0.00	0/25
Moderate	€30	1.0000	—	0.00	0/25
Below 2°C	**€85**	**0.9969**	**< 0.001**	**0.24**	**5/25**
Net Zero	€150	0.9838	< 0.001	0.80	12/25

Note. Carbon prices are reported in euros per tonne of CO2. AADR = Average Absolute Difference in Ranking. The re-ranked column reports projects whose rank changed relative to the baseline. Source: Author calculations.

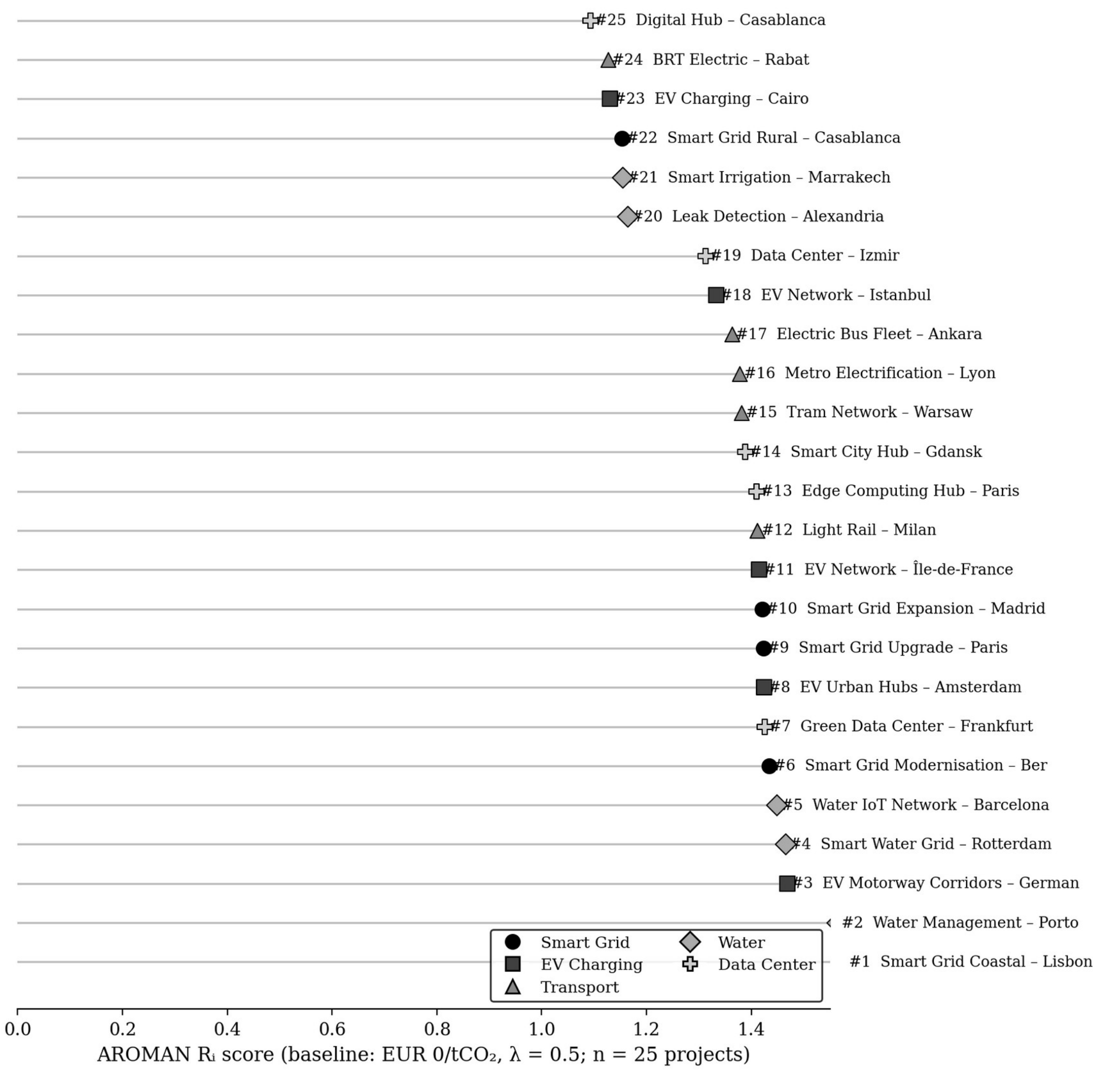

Figure 3. Baseline AROMAN ranking for the 25 smart-city infrastructure projects. Points show final AROMAN scores under the zero-carbon baseline (€0/tCO2) with lambda = 0.5 and corrected two-step normalization (n = 25). Symbols distinguish project types. Higher AROMAN scores indicate stronger multi-criteria investment performance.

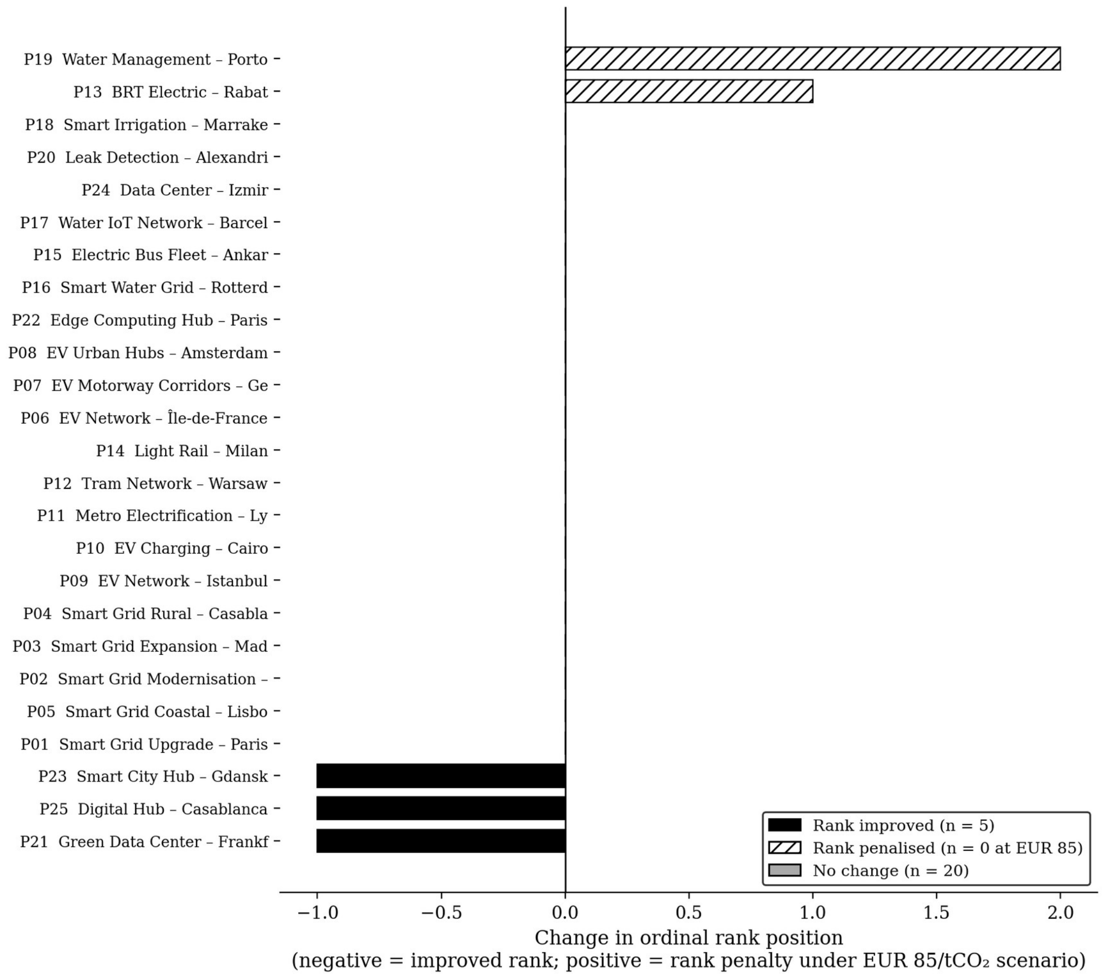

Figure 4. Project-level rank changes under the Below 2°C carbon-price scenario. Bars show differences between the baseline AROMAN rank and the €85/tCO2 scenario rank for the 25-project sample (n = 25). Negative values indicate rank improvement, and positive values indicate rank deterioration. Five projects changed rank under this scenario.

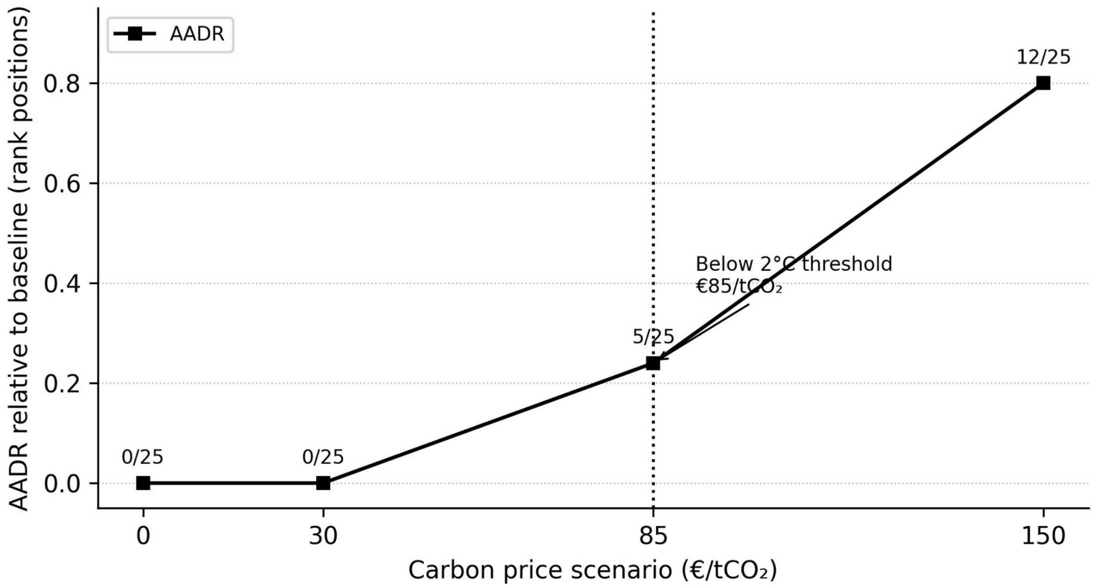

Figure 5. Carbon-price threshold pattern in AROMAN rank stability. Markers show average absolute difference in ranking (AADR, rank positions) for the same 25 projects under Moderate (€30/tCO2), Below 2°C (€85/tCO2), and Net Zero (€150/tCO2) scenarios relative to baseline. The line connects repeated scenario evaluations of the same project set; no independent observations are connected. The €85/tCO2 threshold produced limited but detectable rank disruption, while the €150/tCO2 scenario produced larger ranking changes.

4.3 Hypothesis 3: Geoeconomic Risk and Investment Scores

To test H3, a composite risk-adjusted investment score was constructed for each country-year observation by normalizing and weighting six macro-environmental indicators: FDI inflows, renewable energy consumption, GDP growth (positive drivers), and inflation, CO_2 intensity, and country-level GPR exposure (negative drivers, inverted). An OLS panel regression was then estimated with the composite investment score as the dependent variable and the GPR country index as the key independent variable of interest, controlling for GDP growth, inflation, renewable energy consumption, CO_2 per capita, and country fixed effects.

OLS regression in levels (β = +0.000353, p = 0.928) and OLS in first differences (β = +0.002516, p = 0.500) did not detect a statistically significant negative relationship between the

GPR index and risk-adjusted investment scores at the full panel level (Table 3; Figs. 6 and 7). These results were consistent across model variations and reflected the absence of a detectable panel-level signal, attributable to low temporal GPR variance in European economies and limited MENA observations. Egypt provided the only country-specific confirmation ($\rho = -0.789$, $p < 0.001$), and the GARCH-MIDAS specification was significant at $p = 0.002$. H3 was therefore not confirmed at the panel level and was supported only in the Egypt-specific analysis.

Country-level analysis revealed important heterogeneity. Egypt was the only country exhibiting a statistically significant within-country GPR-investment score relationship (Spearman $\rho = -0.789$, $p < 0.001$, $n = 18$). The GARCH-MIDAS specification for Egypt confirmed that changes in the GPR index predicted FDI return volatility ($\beta = +1{,}234.97$, $p = 0.002$), consistent with Nasouri (2025), who documented stronger GPR effects on financial markets in emerging economies. H3 was not supported at the panel level. Egypt provided exploratory country-specific evidence, but this result could not be generalized to the full sample without a larger and more geographically balanced panel.

Table 3. Geopolitical Risk Tests for AROMAN Score Variation under H3.

Specification	Observations (n)	GPR coefficient (β)	t-statistic	p-value	Verdict
OLS levels — full panel (10 countries)	190	+0.000353	+0.091	0.928	Not significant
OLS first differences — full panel	180	+0.002516	+0.675	0.500	Not significant
GARCH-MIDAS — Egypt (country-specific)	17	+1,234.97	+3.51	**0.002*** **	**Confirmed**
Within-country Spearman ρ — Egypt	18	ρ = −0.789	—	**< 0.001*** **	**Confirmed**
Regime analysis — full panel	190	−0.40 pp avg	—	—	Directional only

Note. OLS estimates used HC3 heteroskedasticity-consistent standard errors. H3 was not supported at the panel level; the Egypt result was interpreted as exploratory country-specific evidence. Source: Author calculations.

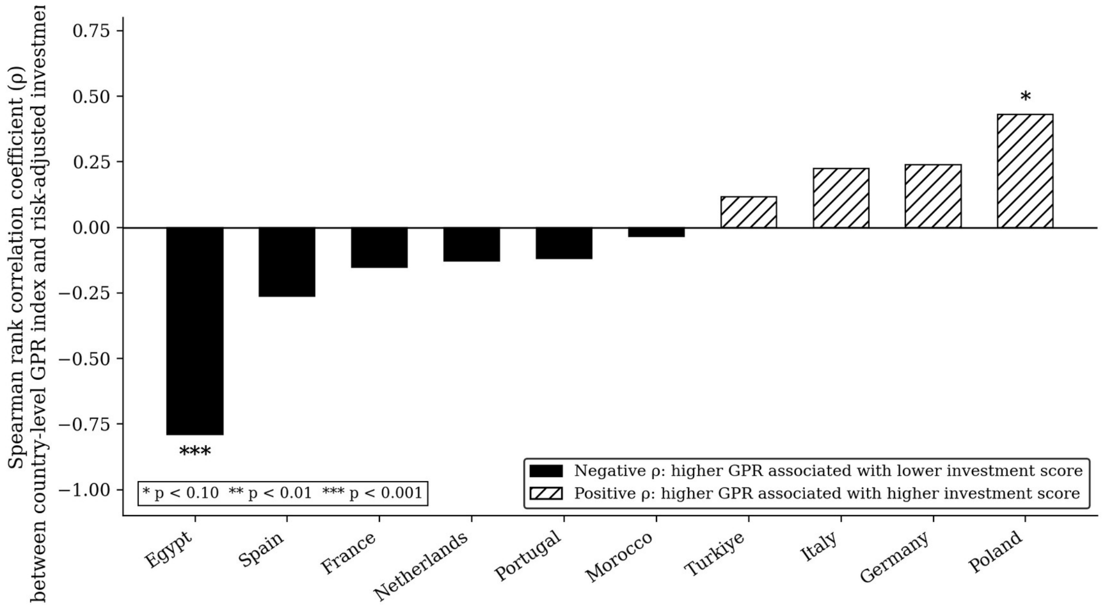

Figure 6. Within-country association between GPR index and risk-adjusted investment score. Bars show Spearman rank-correlation coefficients for ten countries using annual observations from 2003-2021 (n = 18-19 per country, depending on data availability). Negative coefficients indicate lower investment scores in years with higher geopolitical risk. Asterisks denote statistical significance (* p < 0.10, ** p < 0.01, *** p < 0.001).

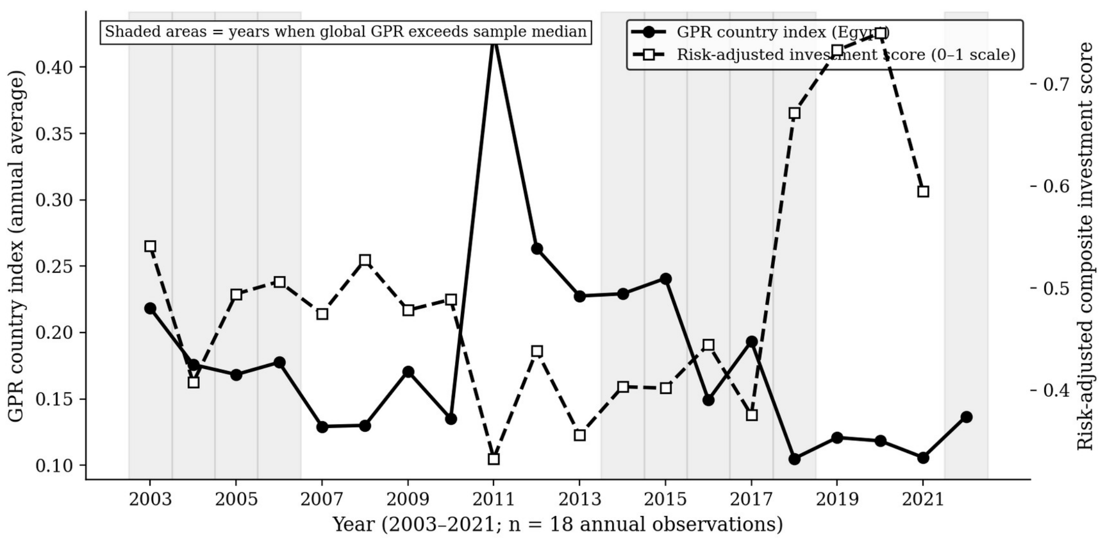

Figure 7. Egypt country-specific GPR pattern and risk-adjusted investment score, 2003-2021. The solid line shows the Egypt GPR country index, and the dashed line shows the risk-adjusted investment score rescaled to 0-1 (n = 18 annual observations). Shaded bands mark years above the sample median of global GPR. The within-country correlation was negative and statistically significant (Spearman rho = -0.789, p < 0.001).

4.4 Hypothesis 4: Model Auditability and Ranking Robustness

H4 was evaluated through four complementary tests: AHP Consistency Ratio (CR) validation of expert-input coherence, One-At-a-Time (OAT) sensitivity analysis tracing rank-change thresholds, a direct comparison of AROMAN multi-criteria rankings against NPV-only rankings, and cross-method concordance with TOPSIS and COPRAS. These tests assessed whether multi-criteria scorecards improved auditability, traceability, and sensitivity disclosure in infrastructure investment rankings compared with single-indicator NPV models.

4.4.1 AHP Consistency Ratio

Consistency Ratio values for all four pairwise comparison matrices satisfied the Saaty (1977) threshold of CR ≤ 0.10, with values ranging from 0.0079 to 0.0190 (Table 4). This result indicated that the criteria hierarchy was internally coherent and that stakeholder-derived weights were not random.

Table 4. AHP Consistency Ratio Results for the Four Pairwise Comparison Matrices.

AHP pairwise comparison matrix	Criteria (n)	lambda_max	Consistency ratio (CR)
Level 1 — Criteria groups	4	4.0514	0.0190
Level 2 — Financial sub-criteria	3	3.0092	0.0079
Level 2 — Environmental sub-criteria	4	4.0514	0.0190
Level 2 — Geoeconomic sub-criteria	3	3.0092	0.0079

Note. lambda_max = principal eigenvalue. CI = Consistency Index. CR = Consistency Ratio. All matrices satisfied the CR < 0.10 threshold; the repeated assessment column was removed to avoid redundant information. Source: Author calculations.

4.4.2 One-at-a-Time Sensitivity Analysis

The OAT analysis systematically varied each of the 13 criteria weights from 0.20 to 1.00 in steps of 0.20, proportionally adjusting remaining criteria to satisfy the unit-sum constraint. The rank-change threshold was uneven across criteria. The most sensitive criterion was NPV, which triggered rank changes when reduced to 0.20, while CO_2 reduction, GPR exposure, and data sovereignty risk triggered rank changes at 0.40. This pattern showed that financial and geoeconomic criteria had greater marginal influence over final rankings than social criteria such as citizen satisfaction.

4.4.3 AROMAN versus NPV-Only Comparison

All 25 projects received different ordinal positions under AROMAN versus the pure NPV-descending ranking in the final corrected ranking file (Spearman $\rho = 0.426$, AADR = 6.40) (Fig. 9). Metro Electrification—Lyon (P11) fell from rank 1 under NPV-only to rank 16 under AROMAN, illustrating how high NPV alone can mask weaker multi-criteria performance. EV Motorway Corridors—Germany (P07) advanced from rank 16 under NPV-only to rank 3 under AROMAN, reflecting stronger multi-criteria performance despite a lower NPV-only position. These divergences illustrate public-interest trade-offs that NPV-only models can suppress.

4.4.4 Cross-Method MCDM Robustness

The cross-method comparison showed high concordance between AROMAN and TOPSIS ($\rho = 0.798$, $p < 0.001$), and between AROMAN and COPRAS ($\rho = 0.739$, $p < 0.001$), when the correct two-step normalization was applied (Table 5; Fig. 8). The results indicated that the scorecard was methodologically robust, while still producing method-specific differences for large-CAPEX projects such as Metro Electrification-Lyon (P11).

Table 5. Cross-Method Ranking Robustness across AROMAN, TOPSIS, and COPRAS.

Project (selected)	AROMAN rank	TOPSIS rank	COPRAS rank	Interpretation
Smart Grid Coastal – Lisbon	**1**	**1**	3	Stable top-3 across all methods
Water Management – Porto	**2**	**2**	**2**	Perfectly stable across all three methods
Metro Electrification – Lyon	16	21	8	Divergence: AROMAN rewards absolute benefit size; TOPSIS rewards benefit–cost efficiency ratio
Spearman ρ vs. AROMAN baseline	**1.000**	**0.798*** **	**0.739*** **	**Strong concordance confirmed with correct two-step normalization (ρ > 0.70)**

Note. MCDM = Multi-Criteria Decision-Making. AROMAN rankings used the corrected two-step normalization procedure. All methods used identical criteria scores and weights. Source: Author calculations.

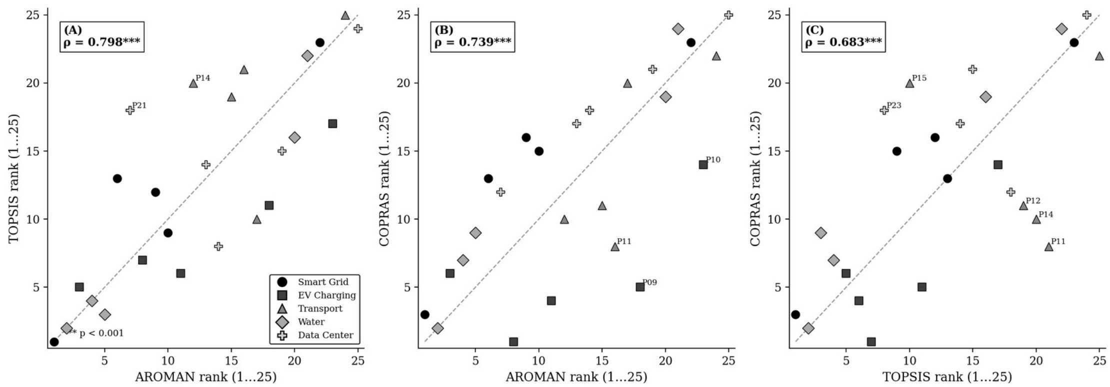

Figure 8. Cross-method robustness of the corrected two-step AROMAN ranking. Scatterplots compare AROMAN ranks with TOPSIS ranks and COPRAS ranks for 25 projects (n = 25). The dotted diagonal indicates perfect rank agreement. Spearman tests showed high concordance between AROMAN and TOPSIS (rho = 0.798, $p < 0.001$) and between AROMAN and COPRAS (rho = 0.739, $p < 0.001$).

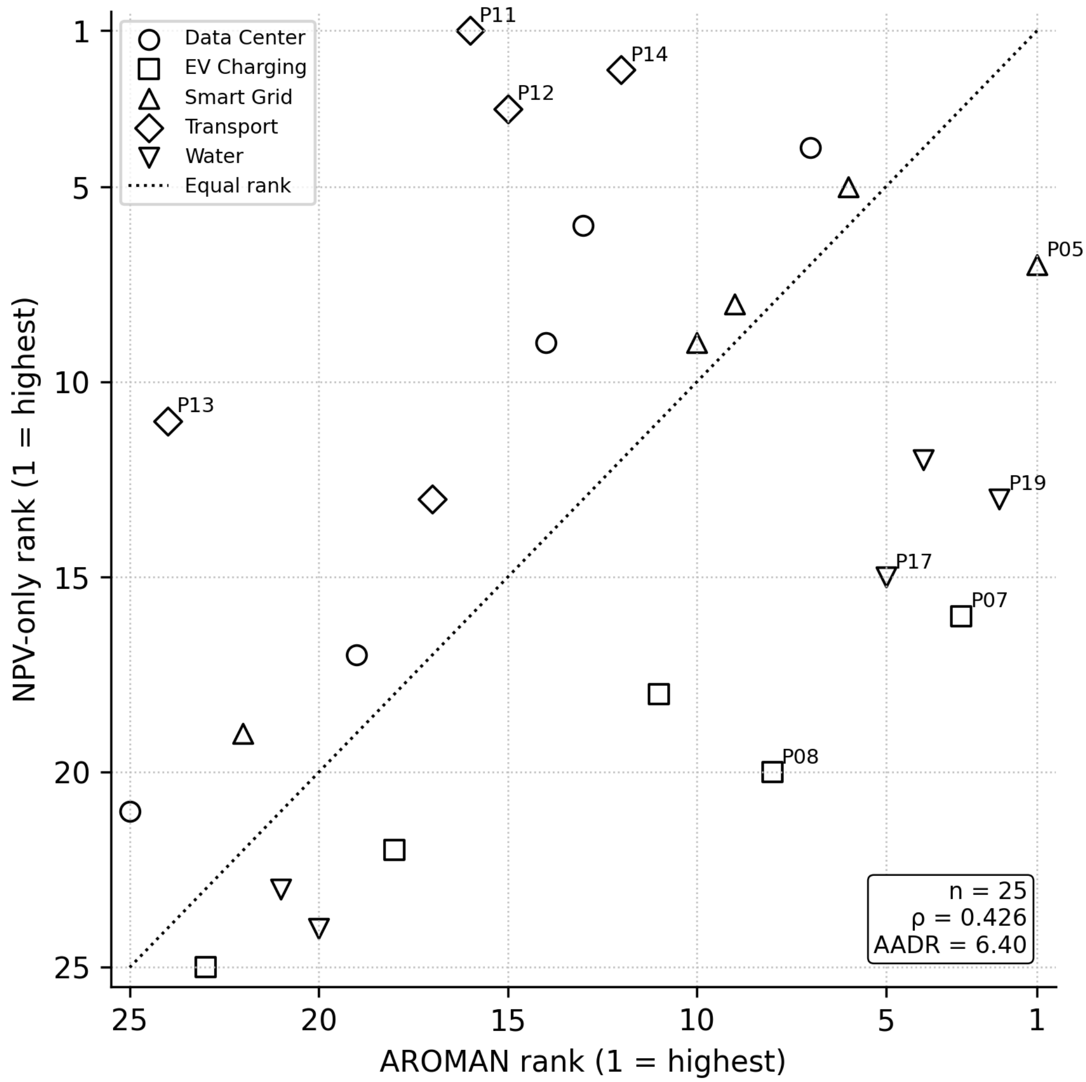

Figure 9. AROMAN multi-criteria ranking compared with NPV-only ranking. Each marker represents one smart-city infrastructure project (n = 25), with marker shape indicating project type. The dotted diagonal shows equal rank positions. All projects received different ordinal positions across the two ranking systems (Spearman rho = 0.426; AADR = 6.40), showing that multi-criteria scoring changed the investment priority order.

4.4.5 Lambda Trade-off Parameter Sensitivity

The λ sensitivity analysis showed high sensitivity at the extremes, with 20-21 projects re-ranked at λ = 0.3 and λ = 0.7, and moderate sensitivity around the baseline (Table 6; Fig. 10). Only P05 and P19 remained stable in the top-2 across the full λ range, confirming that most rankings depended on the normative balance between cost minimization and benefit maximization.

Table 6. Lambda Sensitivity of AROMAN Rankings Relative to the Baseline.

λ value	ρ vs. λ = 0.5 (baseline)	Projects re-ranked	Stable top-5 projects	Interpretation
0.3 (cost-heavy)	0.773	20/25	P05, P19	High sensitivity; cost minimization dominates
0.4	0.892	19/25	P05, P19, P21	Moderate sensitivity
0.5 (balanced — baseline)	**1.000**	**—**	**P05, P19, P07, P16, P17**	**Baseline; equal weight to costs and benefits**
0.6	0.960	18/25	P05, P19, P07	Low sensitivity
0.7 (benefit-heavy)	0.895	21/25	P05, P19	High sensitivity; benefit maximization dominates

Note. lambda is the AROMAN trade-off parameter. Spearman rho compares each lambda value with the lambda = 0.5 baseline. Source: Author calculations.

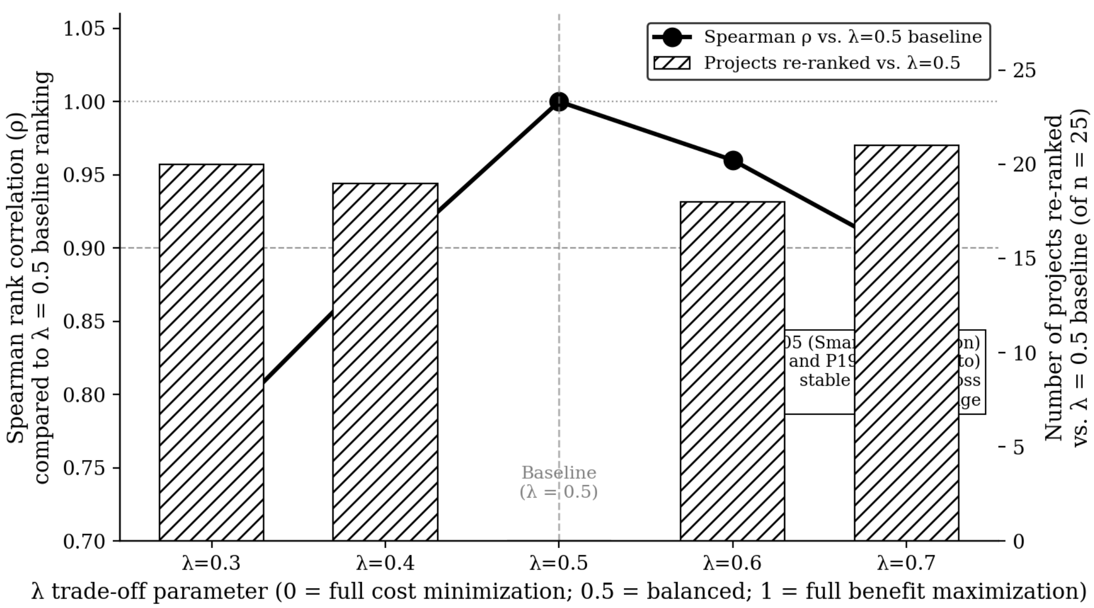

Figure 10. Lambda sensitivity of AROMAN rankings. Bars show the number of projects re-ranked relative to the lambda = 0.5 baseline, and the line shows Spearman rank correlation with the baseline across lambda values from 0.3 to 0.7. The same 25 projects were re-evaluated at each lambda level. Only P05 and P19 remained stable in the top-2 across the full lambda range.

4.5 Summary of Hypothesis Results

Taken together, these results showed that the proposed AI-MCDM framework achieved its core technical objectives: machine learning outperformed deterministic baselines in infrastructure demand forecasting (H1), carbon scenario integration produced limited but detectable ranking effects at €85/tCO_2 and stronger effects at €150/tCO_2 (H2), geoeconomic risk was not supported at the panel level but showed exploratory country-specific evidence for Egypt (H3), and the multi-criteria scorecard supported greater model auditability and ranking traceability than single-indicator financial models (H4).

The hypothesis results were summarized across all four tests (Table 7). H1 and H2 were confirmed. H3 was not confirmed at the panel level, although Egypt provided exploratory country-specific evidence. H4 was supported for model auditability, ranking traceability, and sensitivity disclosure, but not for direct stakeholder consensus or institutional adoption.

Table 7. Hypothesis Results and Supporting Evidence.

Code	Hypothesis	Final status	Key evidence
H1	ML models outperform linear baselines	**Confirmed**	W = 36.00, p = 0.0039; XGBoost −45.8% MAE vs. Ridge; GPR ranked third in feature importance
H2	Carbon scenarios change project rankings	**Confirmed**	ρ = 0.9969, AADR = 0.24; 5/25 re-ranked at €85/tCO_2; 12/25 re-ranked at €150/tCO_2
H3	GPR reduces risk-adjusted investment scores	Not confirmed at panel level	Egypt: ρ = −0.789, p < 0.001; GARCH-MIDAS p = 0.002. Full-panel OLS and first differences not significant
H4	AROMAN scorecard improves auditability and ranking traceability vs. NPV-only	Supported for auditability only	All CR ≤ 0.019; AROMAN–TOPSIS ρ = 0.798***; AROMAN–COPRAS ρ = 0.739***; 25/25 projects ranked differently from NPV-only; AADR = 6.40

Note. pp = percentage points. AADR = Average Absolute Difference in Ranking. H3 was not confirmed at the panel level. Source: Author synthesis from reported tests.

4.6 External Prediction Validation and Replication Outputs

The supplementary WDI/ESG validation panel contained 3,547 usable country-year observations after indicator harmonization and missing-value filtering. The temporal holdout test for 2018-2023 contained 887 observations. XGBoost achieved MAE = 1,608.40 kWh/capita, RMSE = 3,775.00, MAPE = 52.44%, R^2 = 0.587, and Spearman ρ = 0.927. Ridge regression achieved MAE = 1,677.59 kWh/capita, RMSE = 3,818.19, MAPE = 52.54%, R^2 = 0.577, and Spearman ρ = 0.914. These results showed that XGBoost slightly outperformed the linear benchmark on the external country-year panel, although absolute forecasting error remained substantial.

The leave-one-country-out validation evaluated 149 countries. This test assessed geographic transferability rather than project-level ranking validity. The results supported partial external prediction validation of the demand-forecasting component, while the PPI/EIB benchmark supported external plausibility of investment scale and sectoral coverage. The exercise therefore strengthened the paper without converting the AROMAN project rankings into an externally validated prescriptive model.

5. DISCUSSION

The findings reposition smart-city infrastructure appraisal as a multi-risk capital allocation problem instead of a single-indicator financial ranking exercise.

5.1 Theoretical Implications

5.1.1 Integrating MCDM and Dynamic Risk Modeling

The main theoretical contribution lies in the integration of two methodological traditions that prior studies often treated separately. MCDM frameworks such as those used by Pujadas et al. (2017) and Marcelo et al. (2015) support heterogeneous infrastructure prioritization, but they usually rely on static inputs. Climate-finance models such as those developed by In, Weyant, and Manav (2021) and Richstein and Neuhoff (2022) simulate asset-level cash flows under uncertainty, but they rarely embed those outputs in a broader ranking architecture. This study connects these traditions by translating carbon-adjusted DNPV, Real Options outputs, and news-based geoeconomic risk proxies into a single AROMAN decision matrix without removing criterion-level interpretability.

The H1 result extends AI use in infrastructure appraisal beyond cost estimation. XGBoost reduced MAE by 45.8 percent relative to Ridge regression in the original temporal validation, and the external WDI/ESG panel showed a smaller but consistent performance advantage. CO2 per capita, country encoding, and the GPR country index ranked among the largest fitted predictors, which indicates predictive contribution within the model rather than causal influence. This finding supports the need identified by Goloshchapova et al. (2023) and Jin et al. (2023) to include climate, energy, and geoeconomic variables in infrastructure investment models.

5.1.2 Carbon Pricing as a Ranking Disruptor

The H2 findings identify a carbon-price threshold effect in smart-city infrastructure appraisal. Rankings remained unchanged at EUR 30/tCO2, changed slightly at EUR 85/tCO2, and shifted more strongly under the EUR 150/tCO2 Net Zero scenario. Below the threshold, NPV margins absorbed transition costs without changing relative project positions. At higher carbon prices, projects with weaker carbon-exposure-to-NPV ratios experienced rank erosion. This result shows why procurement authorities and infrastructure investors should disclose carbon-price scenarios before treating carbon exposure as a decisive ranking driver.

5.1.3 Geoeconomic Risk as a Heterogeneous Variable

The H3 result limits the theoretical claim about geoeconomic risk. The full panel did not confirm a negative GPR-investment-score relationship, mainly because nine of the ten countries were low-GPR European economies with limited time variation in country-specific GPR series. Egypt produced exploratory country-specific evidence consistent with a negative relationship, and the GARCH-MIDAS specification supported a volatility channel. This pattern aligns with Nasouri (2025), who found stronger GPR transmission in emerging markets. The geoeconomic risk layer therefore appears most useful when portfolios span jurisdictions with heterogeneous geopolitical exposure.

5.1.4 Normative Sensitivity in Multi-Criteria Ranking

The lambda sensitivity analysis adds a theoretical boundary to the scorecard. The trade-off between cost minimization and benefit maximization is not a technical constant; it is a normative parameter with material ranking effects. Only Smart Grid Coastal-Lisbon (P05) and Water Management-Porto (P19) remained stable in the top two across the full lambda range. This finding strengthens the auditability claim under H4 because the framework does not hide stakeholder trade-offs. It exposes them for calibration.

5.2 Managerial and Policy Implications

The AROMAN comparison with NPV-only ranking provides direct evidence that single-indicator financial models can distort infrastructure allocation decisions. In the final corrected ranking file, all 25 projects received different ordinal positions under AROMAN and NPV-only ranking (Spearman $\rho = 0.426$; AADR = 6.40). Metro Electrification—Lyon (P11) fell from rank 1 under NPV-only to rank 16 under AROMAN, while EV Motorway Corridors—Germany (P07) advanced from rank 16 under NPV-only to rank 3 under AROMAN. These divergences show why public authorities should disclose the multi-criteria logic behind capital allocation instead of relying exclusively on NPV.

The carbon scenario analysis provides actionable guidance for development finance institutions and infrastructure debt investors navigating the transition to EU taxonomy-aligned portfolios. The result that Data Center and Smart Grid projects demonstrated greater resilience under Below 2°C and Net Zero scenarios—due to their higher NPV buffers and lower direct ETS exposure—suggests a strategic preference for technology-intensive infrastructure in high-carbon-price environments. Conversely, the framework's GPR layer adds information that is absent from conventional ESG scoring systems: the data sovereignty risk dimension, operationalized through a qualitative NIS2/GDPR compliance cost proxy, penalizes projects in jurisdictions with weak digital governance frameworks. This dimension is increasingly relevant to smart city data center investment decisions during the phased implementation of the EU Cyber Resilience Act and CBAM's shift from transitional reporting to its definitive regime (European Commission, 2025, 2026).

From a research positioning perspective, the framework's explicit integration of supply-delay risk, imported-technology dependence, and data-sovereignty compliance cost as geoeconomic sub-criteria represents a novel operationalization of the geopolitical risk concept in smart city investment evaluation. Prior MCDM studies in this domain — including Angelakoglou et al. (POCITYF, 2022) and the Konya smart city AHP study — do not include geoeconomic variables in their criteria hierarchies. The present framework therefore offers a methodological template for extending ISO 37120-aligned smart city evaluation frameworks to incorporate the geopolitical and regulatory risk dimensions that the post-pandemic, post-Ukraine investment environment now demands.

Managerial takeaway. The framework does not replace financial appraisal; it complements financial appraisal by identifying when projects with moderate NPV may generate stronger public value because of carbon resilience, geoeconomic exposure, or strategic infrastructure benefits.

5.3 Robustness and Reporting Boundaries

The evidence supports proof-of-concept validation of an AI-enabled MCDA scorecard, not a definitive project-selection model. The framework integrates financial, carbon, energy, and geoeconomic indicators into a traceable ranking structure. The evidence supports model auditability, ranking traceability, and sensitivity disclosure, not stakeholder consensus or institutional adoption. Audited project-level datasets remain necessary before the framework can support prescriptive investment decisions (Collins et al., 2024; Moons et al., 2025).

AROMAN produces a structured decision-support ranking, not an objective optimization result. Final ranks depend on criterion selection, weighting assumptions, normalization logic, and the λ trade-off parameter. This boundary condition shapes H4: the framework improves ranking auditability but does not replace expert judgment, institutional deliberation, or project-level due diligence.

An external benchmarking pass was added to reduce reliance on the constructed 25-project sample. The benchmark used the World Bank PPI 2024 full project file, EIB financed-project records, EIB pipeline projects, EIBIS indicators, and the EIB Municipalities Survey. The World Bank PPI benchmark contained 8,213 project observations for 2000-2024, covering 129 countries and all five infrastructure sectors used in this study: energy, transport, water and sewerage, ICT, and municipal solid waste. Of these observations, 7,606 included investment values. The EIB financed-project benchmark contained 7,619 relevant infrastructure loan records signed between 2000 and 2024 across 155 countries, with a total signed amount of approximately EUR 742.2 billion. A keyword-filtered EIB smart-infrastructure subset contained 3,894 records across 131 countries and EUR 388.2 billion in signed financing.

The external benchmark supports scale plausibility, not full predictive external validity. The constructed project sample had an initial investment range of EUR 41-320 million and a median of EUR 118 million. This range falls within the observed PPI and EIB infrastructure-finance distributions: the PPI median investment was USD 77 million overall, with sector medians of USD 75 million for energy, USD 156.5 million for transport, USD 30.4 million for water, USD 55 million for data-center projects, and USD 142.5 million for electric-vehicle charging projects. The EIB relevant infrastructure-loan median was EUR 50 million. Therefore, the benchmark indicates that the sample is financially plausible relative to real infrastructure portfolios, while the absence of audited project-level outcomes still prevents a complete external validation of forecasting accuracy or final investment rankings.

5.4 Limitations

Despite these contributions and the added external benchmarking pass, several methodological and empirical limitations continue to reduce the generalizability and robustness of the findings. These limitations are summarized in Table 8 and should guide future work.

First, the project sample and key performance parameters were externally benchmarked against World Bank PPI and EIB project records, but they were not replaced with independently audited project-level outcome data. The benchmarking pass supports financial scale plausibility because the constructed projects fall within observed PPI and EIB investment distributions. However, the absence of verified carbon footprints, real-time EU ETS exposure assessments, post-completion satisfaction metrics, and out-of-sample demand outcomes still limits external validity. Independent external data are necessary for unbiased assessment of predictive model performance and generalizability (Gallitto et al., 2025). Future versions should therefore include leave-one-country-out validation, external validation on an independent project dataset, and uncertainty intervals around XGBoost forecasts.

Second, the weighting process relied on AHP judgments that assumed independence among criteria, used a fixed 1-9 comparison scale, and provided a static snapshot of priorities. These assumptions were appropriate for a proof-of-concept scorecard but remain restrictive for complex infrastructure portfolios, where criteria may interact and expert judgments may be inconsistent (Belton & Gear, 1983; Forman & Gass, 2001; Frish et al., 2025). Future applications should co-design weights with stakeholders, report consistency ratios, test weight perturbations, and disclose the normative implications of weighting choices.

Third, the data sovereignty risk criterion - currently operationalized as a qualitative 1-5 ordinal score - should be replaced with a quantitative composite index constructed from NIS2 compliance cost estimates, GDPR enforcement action frequency, cloud localization requirements, and OECD digital services trade restriction indices. This operationalization would increase the precision and replicability of the geoeconomic risk layer and reduce the AHP subjectivity that constrains H4's generalizability.

Fourth, the XGBoost-based demand forecasting and feature-importance analysis provided limited interpretability. Studies show that tree-based models such as XGBoost, with or without SHAP, may assign importance to irrelevant features, which limits the evidential strength of feature-importance claims (Dunn et al., 2021). Feature-importance outputs should therefore be interpreted as variables contributing to predictive performance within the fitted model, not as causal explanations of infrastructure demand. Future work should complement SHAP with permutation importance, partial dependence plots, and simpler benchmark models to separate predictive contribution from causal interpretation.

Fifth, the sample exhibited geographic imbalance: nine of ten countries were low-GPR European economies, which compressed geoeconomic risk variation and reduced the statistical power of H3. Expanding the dataset to include more MENA, Sub-Saharan African, and South-East Asian economies, and incorporating country-specific GPR series for Morocco, Saudi Arabia, and Vietnam, would provide a more rigorous test of the hypothesised transmission channel.

Table 8. Principal Methodological Limitations and Proposed Mitigations.

Limitation	Nature	Proposed mitigation
Benchmarked 25-project sample; no audited outcomes	External validity	Use PPI/EIB benchmarking as plausibility evidence; validate next against audited post-completion data
External benchmark added, full external validation still incomplete	Predictive generalizability	Freeze the workflow; run leave-one-country-out and independent project-level validation
AHP weights depend on expert judgments	Weighting subjectivity	Use structured elicitation, report CR values, and run weight sensitivity tests
Data sovereignty risk uses ordinal 1-5 scoring	Measurement	Replace with a quantitative composite index using NIS2, GDPR, and OECD STRI inputs
XGBoost feature importance is predictive, not causal	Interpretability	Use SHAP, permutation importance, PDPs, and simpler benchmark models together
H3 underpowered by low-GPR European sample	Statistical power	Expand to MENA, Sub-Saharan Africa, and South-East Asia; add monthly GPR series

Note. STRI = OECD Services Trade Restrictiveness Index. NIS2 = EU Network and Information Security Directive. The limitations define the valid interpretation boundaries of the proof-of-concept scorecard. Source: Author synthesis.

5.5 Directions for Future Research

The preceding discussion shows that the proposed AI-MCDM framework makes three original contributions to the infrastructure investment evaluation literature. First, it integrates

XGBoost-based demand forecasting, carbon-adjusted DNPV valuation, and news-based geoeconomic risk proxies within a single AROMAN decision matrix, closing the integration gap identified by Goloshchapova et al. (2023) and Jin et al. (2023). Second, it identifies a carbon price threshold effect at approximately €85/tCO_2, although the ranking impact becomes stronger under the €150/tCO_2 Net Zero scenario. Third, it shows that multi-criteria scoring can produce rankings that diverge from NPV-only evaluation, providing quantitative evidence that single-indicator models may misallocate capital in heterogeneous public infrastructure portfolios. These contributions position the framework as a methodological advance and a practical decision-support tool for institutional actors navigating smart city investment under uncertainty.

Five directions for future research emerge directly from the limitations and findings of this study.

First, the forecasting component should be validated through leave-one-country-out testing, external project-level validation, and prediction-interval reporting. These procedures would show whether the model generalizes beyond country-specific patterns and would align the AI component more closely with TRIPOD+AI and PROBAST+AI expectations for transparent predictive-model reporting and bias assessment (Collins et al., 2024; Moons et al., 2025).

Second, the GPR panel regression should be extended to a larger cross-country sample incorporating MENA, Sub-Saharan African, and South-East Asian economies, where geoeconomic risk variance is higher and the hypothesised transmission mechanism can be tested with greater statistical power. The inclusion of country-specific GPR series for Morocco, Saudi Arabia, and Vietnam—jurisdictions actively developing smart city infrastructure programs—would improve H3's empirical testability.

Third, the GARCH-MIDAS specification should be re-estimated using monthly energy price data from the EIA and Nordpool, combined with the monthly GPR index, to capture the high-frequency volatility dynamics that the annual panel dataset cannot resolve. This extension would align the empirical methodology more closely with the theoretical model of Nasouri (2025) and provide a more rigorous test of the GPR-to-volatility transmission channel.

Fourth, the data sovereignty risk criterion - currently operationalized as a qualitative 1-5 ordinal score - should be replaced with a quantitative composite index constructed from NIS2 compliance cost estimates, GDPR enforcement action frequency, cloud localization requirements, and OECD digital services trade restriction indices. This operationalization would increase the precision and replicability of the geoeconomic risk layer and reduce the AHP subjectivity that constrains H4's generalizability.

Fifth, the framework should be extended to incorporate a real-time dashboard interface that dynamically updates AROMAN rankings as new GPR index values, EU ETS prices, and project performance data are published. The Python/Dash stack used in this research provides a viable technical foundation for such an implementation, which would transform the academic scorecard into an operational decision-support tool for urban investment authorities.

6. CONCLUSIONS

This paper answers a single integrative research question: how can an AI-enabled multi-criteria investment scorecard integrate financial indicators, energy-price scenarios, carbon-cost exposure, and geoeconomic-risk proxies to rank smart city infrastructure projects with greater auditability and risk sensitivity than traditional financial-only models? The empirical and methodological results presented across Sections 3 and 4 provide a structured answer.

The proposed four-layer framework translates GPR index signals and EU ETS carbon price trajectories through asset-level DNPV and Real Options valuation, aggregates outputs via Fuzzy-AHP and MEREC weighting, and produces final rankings through the AROMAN normalization formula. H1 and H2 are confirmed; H3 is not confirmed at the panel level but produces exploratory Egypt-specific evidence; and H4 is supported for model auditability, ranking traceability, and sensitivity disclosure rather than direct stakeholder validation.

H1 establishes that XGBoost achieves a 45.8 percent reduction in MAE relative to Ridge linear regression for infrastructure demand forecasting, with the Wilcoxon test confirming significance at $p = 0.0039$. H2 identifies a carbon price threshold effect: rankings are invariant below €30/tCO_2, limited but detectable at €85/tCO_2, and more disrupted under the €150/tCO_2 Net Zero scenario. H3 is not supported at the panel level; country-specific confirmation for Egypt remains exploratory ($\rho = -0.789$, $p < 0.001$), and future research should expand the panel before generalizing this mechanism. H4 is supported through AHP Consistency Ratios below 0.02 across all matrices, high cross-method concordance with TOPSIS and COPRAS, and a direct comparison showing that AROMAN and NPV-only rankings diverge across all 25 projects in the final corrected ranking file.

The four original contributions of this research are summarized in the contribution matrix (Table 9).

Table 9. Original Contributions and Supporting Evidence.

Contribution code	**Contribution**	**Supporting evidence**
C1	First unified AI-MCDM scorecard integrating GPR, EU ETS carbon costs, and energy price scenarios within a single auditable decision matrix	Four-layer framework: GPR + ETS → DNPV/ROV → Fuzzy-AHP/MEREC → AROMAN with correct two-step normalization (Đalić et al., 2021)
C2	XGBoost with geoeconomic features outperforms linear baselines for infrastructure demand forecasting	W = 36.00, p = 0.0039; 45.8% MAE reduction; GPR country index ranked third in feature importance (H1)
C3	Carbon price threshold effect with limited impact at €85/tCO_2 and stronger disruption under the €150/tCO_2 Net Zero scenario	At €85/tCO_2: ρ = 0.9969, AADR = 0.24; at €150/tCO_2: 12/25 projects re-ranked (H2)
C4	AROMAN scorecard diverges from NPV-only ranking and remains concordant with TOPSIS/COPRAS; λ normative sensitivity is quantified and disclosed	AADR = 6.40; 25/25 projects ranked differently vs. NPV-only; CR ≤ 0.019 across AHP matrices; only P05 and P19 stable across full λ range (H4)

Note. Contributions were classified only when a result, robustness test, or external benchmark supported the claim. Source: Author synthesis.

Smart city infrastructure investment has become a defining capital allocation challenge. The framework does not eliminate uncertainty. No model can do so. Instead, the framework makes uncertainty visible, traceable, and actionable. The scorecard embeds carbon scenario sensitivity, geoeconomic risk exposure, multi-criteria weighting, and explicit λ normative disclosure. This structure supports urban planners, institutional investors, and development banks as they allocate capital under financial, environmental, and geopolitical constraints.

DATA AND CODE AVAILABILITY

The replication package contains cleaned analytical datasets, Python scripts, validation outputs, supplementary materials, figure files, and a data-source inventory. The public source datasets include World Bank WDI/ESG indicators, the World Bank PPI Database, EIB financed-project records, EIB pipeline records, EIBIS indicators, and the EIB Municipalities Survey. The constructed smart-city project dataset, processed WDI/ESG panel, validation scripts, and supplementary tables are deposited in a public repository. The replication package is available at https://doi.org/10.5281/zenodo.20044642.

REFERENCES

Abdelkader, E. M., Zayed, T., & F., H. E. (2023). A two-tier hybrid model for selecting private partners in PPP transportation projects. Mathematics, 11(16), 3559. https://doi.org/10.3390/math11163559

Alizadeh, T. (2017). An investigation of IBM's Smarter Cities Challenge: What do participating cities want? Cities, 63, 70–80. https://doi.org/10.1016/j.cities.2016.12.009

Assab, A. (2024). Theoretical foundation for pricing climate-related loss and damage in infrastructure financing. Journal of Risk and Financial Management, 17(4), 133. https://www.mdpi.com/1911-8074/17/4/133

Belton, V., & Gear, T. (1983). On a shortcoming of Saaty's method of analytic hierarchies. Omega, 11(3), 228-230. https://doi.org/10.1016/0305-0483(83)90047-6

Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. American Economic Review, 112(4), 1194–1225. https://doi.org/10.1257/aer.20191823

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

Cox, L. A., Jr. (2008). What's wrong with risk matrices? Risk Analysis, 28(2), 497-512. https://doi.org/10.1111/j.1539-6924.2008.01030.x

Đalić, I., Stević, Ž., Erceg, Ž., Macura, D., & Terzić, S. (2021). Selection of a distribution channel using the AROMAN method. Decision Making: Applications in Management and Engineering, 4(2), 107–127. https://doi.org/10.31181/dmame210402107d

Dunn, J., Mingardi, L., & Zhuo, Y. D. (2021). Comparing interpretability and explainability for feature selection. arXiv. https://doi.org/10.48550/arXiv.2105.05328

Espinoza, D., Rojo, J., Phillips, W., & Eild, A. (2022). Decoupled net present value. Sustainable and Resilient Infrastructure. https://doi.org/10.1080/23789689.2022.2148453

European Commission. (2025). Cyber Resilience Act. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

European Commission. (2026). Start of the definitive period of the CBAM in the EU. Access2Markets. https://trade.ec.europa.eu/access-to-markets/en/news/start-definitive-period-cbam-eu

European Investment Bank. (2025a). EIB Investment Survey (EIBIS) [Data set]. https://data.eib.org/eibis/download

European Investment Bank. (2025b). Financed projects [Data set]. https://www.eib.org/en/projects/loans/index

European Investment Bank. (2025c). The state of local infrastructure investment in Europe: EIB Municipalities Survey 2024-2025. https://doi.org/10.2867/3529559

Flora, M., & Tankov, P. (2023). Green investment and asset stranding under transition scenario uncertainty. Energy Economics, 124, 106773. https://doi.org/10.1016/j.eneco.2023.106773

Forman, E. H., & Gass, S. I. (2001). The analytic hierarchy process-An exposition. Operations Research, 49(4), 469-486. https://doi.org/10.1287/opre.49.4.469.11231

Frish, S., Talmor, I., Hadar, O., Shoshany, M., & Shapira, A. (2025). Enhancing consistency of AHP-based expert judgements: A new approach and its implementation in an interactive tool. MethodsX, 14, 103341. https://doi.org/10.1016/j.mex.2025.103341

Gallitto, G., Englert, R., Kincses, B., Kotikalapudi, R., Li, J., Hoffschlag, K., & Bingel, U. (2025). External validation of machine learning models-registered models and adaptive sample splitting. GigaScience, 14, giaf036. https://doi.org/10.1093/gigascience/giaf036

Gatti, S. (2024). Project Finance in Theory and Practice (4th ed.). Academic Press. https://doi.org/10.1016/B978-0-323-98360-0.00001-9

Goloshchapova, I., Poulter, R., & Paulson, A. (2023). Geopolitical risk and energy investment: Why quantitative models fall short. Energy Policy, 178, 113583. https://doi.org/10.1016/j.enpol.2023.113583

In, S. Y., Weyant, J. P., & Manav, B. (2021). Pricing climate-related risks of energy investments. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3943174

Jin, Y., Wu, F., & Zhao, C. (2023). The dynamic effects of geopolitical risks on the clean energy transition across different economic development stages. Renewable and Sustainable Energy Reviews, 182, 113345. https://doi.org/10.1016/j.rser.2023.113345

Magwedere, M. R., & Marozva, G. (2023). Does political risk matter for infrastructure investments? Empirical evidence. Cogent Economics & Finance, 11(1), 2146596. https://doi.org/10.1080/23322039.2022.2146596

Marcelo, D., Mandri-Perrott, C., House, S., & Schwartz, J. Z. (2015). Prioritizing infrastructure investment: A framework for government decision making. World Bank Policy Research Working Paper No. 7414. https://doi.org/10.1596/1813-9450-7414

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., et al. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388, e082505. https://doi.org/10.1136/bmj-2024-082505

Narbaev, T., Hazir, Ö., & İnan, T. (2023). A machine learning study to improve project cost estimate reliability. International Journal of Production Research, 62(12), 4372–4388. https://doi.org/10.1080/00207543.2023.2262051

Nasouri, A. (2025). The impact of geopolitical risks on equity markets and financial stress: A comparative analysis of emerging and advanced economies. International Journal of Economics and Business Administration, XIII(1), 30–41. https://www.ijeba.com/journal/873

Pujadas, P., Pardo-Bosch, F., Aguado-Renter, A., & Aguado, A. (2017). MIVES multi-criteria approach for the evaluation, prioritization, and selection of public investment projects: A case study in the city of Barcelona. Land Use Policy, 64, 29–37. https://doi.org/10.1016/j.landusepol.2017.02.014

Rahman Sourav, M. S., Hossain, A., Islam, M. R., Wasif, M., & Samia, S. (2025). AI-driven forecasting in BRICS infrastructure investment: Impacts on resource allocation and project delivery. Journal of Economics, Finance and Accounting Studies, 7(2), 117–132. https://doi.org/10.32996/jefas.2025.7.2.11

Sany, E. G. (2026). Replication Package for AI-Enabled Investment Scorecards for Smart-City Infrastructure Financing (1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.20044642

Richstein, J. C., & Neuhoff, K. (2022). Carbon contracts-for-difference: How to de-risk innovative low-carbon investments. iScience, 25(8), 104700. https://doi.org/10.1016/j.isci.2022.104700

Sarmas, E., Marinakis, V., & Doukas, H. (2022). A data-driven multicriteria decision making tool for assessing investments in energy efficiency. Operational Research, 22(5), 5597–5616. https://doi.org/10.1007/s12351-022-00727-9

Shamim, M. M. I., Hamid, A. B. b. A., Nyamasvisva, T. E., & Rafi, N. S. B. (2025). Advancement of artificial intelligence in cost estimation for project management. Modeling, 6(2), 35. https://doi.org/10.3390/modeling6020035

Sing, M. C. P., Ma, Q., & Gu, Q. (2025). Improving cost contingency estimation with artificial neural networks and a complexity index. Applied Sciences, 15(7), 3519. https://doi.org/10.3390/app15073519

Su, C.-W., Liu, F., Qin, M., & Bouri, E. (2023). Are climate and geopolitics the challenges to sustainable development? Novel evidence from the global supply chain. Economic Analysis and Policy, 77, 748–763. https://doi.org/10.1016/j.eap.2023.01.002

World Bank. (2024). Private Participation in Infrastructure (PPI) Database: Project all data [Data set]. https://ppi.worldbank.org/en/ppidata

APPENDIX A: CRITERIA OPERATIONALIZATION

The exact operationalization of the 13 scoring criteria is documented to ensure full replicability of the project evaluation matrix (Table 10). Financial criteria were benchmarked against EIB project financing summaries scaled to 2022 euros using the Eurostat Harmonised Index of Consumer Prices deflator. Environmental criteria were calibrated against EU POCITYF thresholds and the EIB Greenhouse Gas Assessment Methodology (2023 edition). Social criteria were derived from IESE Cities in Motion Index benchmarks and ILO infrastructure employment multipliers. GPR exposure used Caldara and Iacoviello (2022) country-specific annual averages; Morocco used the global GPR index because no country-specific series was available, a limitation noted in the study (Table 8). Data sovereignty risk used a three-component composite: NIS2 implementation status, GDPR enforcement frequency, and the OECD digital services trade restriction index.

Table 10. Operationalization of the Thirteen Project Evaluation Criteria.

Criterion	Type	Primary source	Operationalization
Initial investment (M€)	Financial	EIB archives	Total CAPEX from EIB project financing summaries; scaled to 2022 euros using Eurostat HICP deflator
Net present value (M€)	Financial	EIB / WB PPI	20-year NPV at 8% social discount rate; revenues from EIB project-type benchmark reports
O&M cost (M€/yr)	Financial	EIB archives	Annual operations and maintenance cost from EIB Technical Advisor reports; typically 3–5% of CAPEX for smart infrastructure
CO_2 reduction (%)	Environmental	EIB GHG (2023)	Estimated annual CO_2 reduction vs. baseline using EIB Greenhouse Gas Assessment Methodology (2023 edition) by project type
Energy savings (%)	Environmental	POCITYF	Reduction in primary energy demand; calibrated against EU POCITYF threshold: 32.5% reduction = ordinal score 3/5
Renewable energy rate (%)	Environmental	WB WDI	Country-level renewable energy consumption as % of total final energy; WDI 2018–2022 average
Access to electricity (%)	Infrastructure	WB WDI	National access to electricity as % of population; WDI 2021. Proxy for digital infrastructure readiness
Citizen satisfaction (1–5)	Social	IESE CIMI / POCITYF	Ordinal score derived from IESE Cities in Motion Index social cohesion dimension and POCITYF Likert-scale survey benchmarks

Criterion	Type	Primary source	Operationalization
Job creation (direct jobs)	Social	ILO / WB	Direct job-years estimated from ILO infrastructure employment multipliers by project type and host-country income level
Technology integration score (1–5)	Smart city	ISO 37122	Ordinal score based on ISO 37122 smart city technological readiness indicators; score 5 = full IoT and AI integration
GPR exposure	Geoeconomic	Caldara & Iacoviello (2022)	Country-specific annual GPR average 2003–2022. Morocco uses global GPR index (no country series available — documented as limitation in Table 8)
Data sovereignty risk (1–5)	Geoeconomic	OECD STRI / NIS2	Three-component composite: NIS2 implementation status + GDPR enforcement frequency + OECD digital services trade restriction index
ETS carbon exposure (M€)	Carbon risk	EMBER / EIB	Annual carbon liability = project direct emissions (tCO_2) × ETS price. Calibrated using EIB carbon footprint coefficients by infrastructure type

Note. WDI = World Bank World Development Indicators. HICP = Harmonised Index of Consumer Prices. STRI = OECD Services Trade Restrictiveness Index. CIMI = Cities in Motion Index. O&M = operations and maintenance. Source: Author synthesis.

APPENDIX B: AI-MCDM REPORTING CHECKLIST

The checklist summarizes the reporting boundaries of the AI-enabled MCDA scorecard by adapting prediction-model reporting principles to the infrastructure investment context (Table 11; Collins et al., 2024; Moons et al., 2025). The checklist does not claim formal TRIPOD+AI compliance; it helps reviewers assess transparency.

Table 11. AI-MCDA Reporting Checklist for Prediction and Ranking Transparency.

Item	Status	Location
Research question	Reported	§1.3
Data sources	Reported	§3.2
Project sample	Reported	§3.3
Predictors and criteria	Reported	§3.3; Appendix A
ML model specification	Reported	§3.4; §4.1
Internal validation	Partial	§4.1
External benchmarking	Performed; full validation future	§3.2; §5.3; Appendix C
Bias and applicability	Discussed	§5.3; §5.4
MCDA weighting	Reported	§3.4; Table 2
Sensitivity analysis	Reported	§4.4
Interpretability limits	Discussed	§5.4
Stakeholder validation	Not performed	§5.3; §5.4

Note. MCDA = Multi-Criteria Decision Analysis. ML = Machine Learning. The checklist adapts prediction-model reporting principles to an AI-enabled MCDA scorecard. Source: Author synthesis.

APPENDIX C: EXTERNAL BENCHMARKING AND TRANSFERABILITY CHECKS

Appendix C reports the external benchmarking pass used to assess whether the constructed 25-project scorecard sample was plausible relative to independent infrastructure-finance records (Table 12). These checks benchmark scale, sectoral coverage, and policy relevance; they do not constitute full external validation of final AROMAN rankings because audited project outcomes and the original demand-forecasting panel are not available in the external sources.

Table 12. External Benchmarking Sources and Transferability Role.

Benchmark source	Usable observations	Coverage	Validation role	Main finding
World Bank PPI full data (2000-2024)	8,213 project observations; 7,606 with investment values	129 countries; energy, transport, water/sewerage, ICT, MSW	International infrastructure-finance benchmark	The sample median CAPEX (EUR 118m) falls within observed PPI investment distributions; PPI median investment = USD 77m.
World Bank PPI by project-type proxy	Energy 4,592; transport 1,800; water 1,392; data center 51; EV charging 24	Low- and middle-income infrastructure portfolios	Sectoral scale benchmark	Project-type medians support plausibility for transport, smart grid/energy, water, data-center and EV-charging categories, with EV charging sparse.
EIB financed projects (2000-2024)	7,619 relevant infrastructure loan records	155 countries; energy, transport, water, telecom, solid waste, urban development	EIB/EU institutional benchmark	Relevant infrastructure loans total approximately EUR 742.2bn; median signed amount = EUR 50m.
EIB smart-infrastructure keyword subset	3,894 records	131 countries; grid, digital, data, EV, water, urban, transport, energy keywords	Thematic plausibility benchmark	The smart-infrastructure subset totals approximately EUR 388.2bn; median signed amount = EUR 50m.
EIB pipeline, EIBIS, and Municipalities Survey	660 relevant pipeline projects; EIBIS 2025 all-sector/all-size indicators; 1,002 municipalities	EIB forward pipeline; EU27/US firm indicators; EU municipalities	Contextual validation	Finance, regulation, climate, digitalisation and trade-disruption indicators support the inclusion of non-financial criteria in the scorecard.

Note. PPI = Private Participation in Infrastructure. EIB = European Investment Bank. EIBIS = EIB Investment Survey. MSW = Municipal Solid Waste. Source: Author calculations from the replication package.