Supplementary Information for "Meeting equity requirements in shared micromobility rebalancing: a constrained Markov decision process with a case study in The Hague"

Lorenzo Rota Canmanie T. Ponnambalam Thiago D. Simão

Journal: Discover Cities. Corresponding author: Lorenzo Rota (l.rota@student.tue.nl).

S1 Additional scenarios

The main text evaluates the five-category scenario; the two-, three-, and four-category scenarios of Cederle et al. [1] were trained and evaluated under the same protocol to check that the behavior generalizes. Figure S1 shows per-category failure rates and rebalancing cost against $r_{\max}$ for each.

The pattern of the five-category case holds in all three. The worst-case morning failure rate decreases from about 25% when unconstrained toward the bound as $r_{\max}$ tightens, the bound is met at every setting, and the rebalancing cost rises as the bound tightens. As in the main text, the evening constraint binds only in the scenarios that contain category 2, since it carries the high evening demand; in the two- and four-category scenarios the evening constraint is slack while the morning constraint drives the strategy. The cost level scales with the number of constrained stations, from 70 in the two-category scenario to 130 in the four-category one.

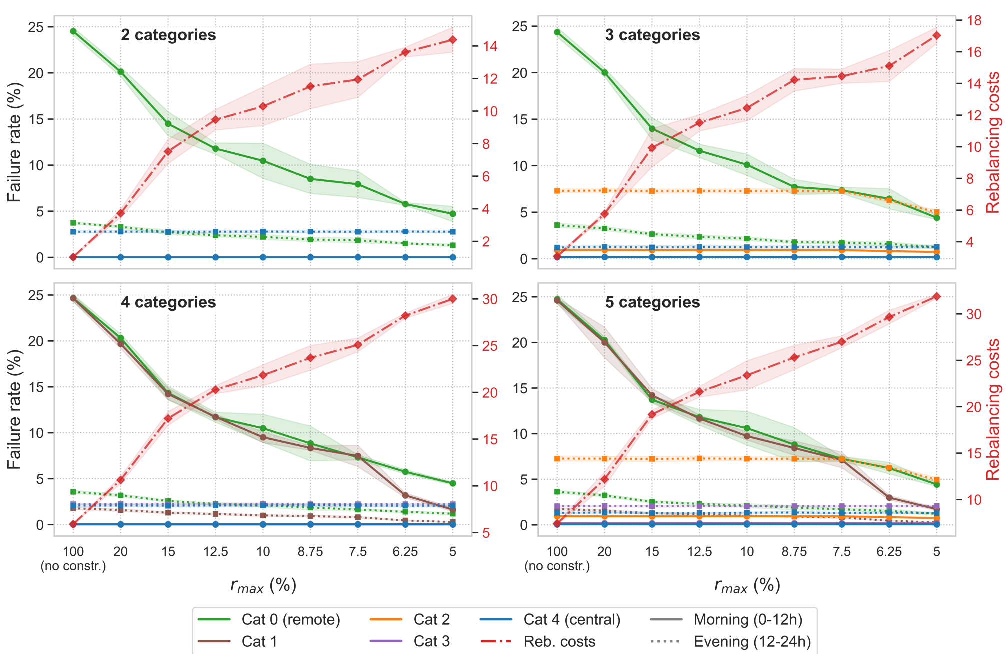

Figure S1: Per-category failure rates (left axis) and weighted rebalancing operations (right axis) against $r_{\max}$ for the two-, three-, four-, and five-category scenarios, over 10 seeds

S2 Dual-variable dynamics

Figure S2 shows, for the synthetic five-category scenario, the dual variable λ for category 0, the most peripheral category and the one whose constraint binds hardest, across the five tightest bounds and separated by period. Each point is one dual update, taken every n_{dual} days. Early in training the morning dual (solid) rises sharply as failures exceed the threshold, peaking higher for tighter bounds, then falls once the policy has cut failures. When the constraint is met it relaxes toward zero, failures drift back up, and it rises again, oscillating around a low value. This is the expected behavior of a fixed-step Lagrangian update, whose dual responds to policy behavior [2]. The evening dual (dashed) stays near zero throughout, consistent with the non-binding evening constraints.

Because the strategy is trained under this oscillating penalty, it holds failures in a band that is on average consistent with the bound rather than resting exactly on it, which is why the achieved failure rate in the main text approaches but does not sit on $r_{\max}$. A damped dual update would reduce the oscillation, as noted in the main-text discussion.

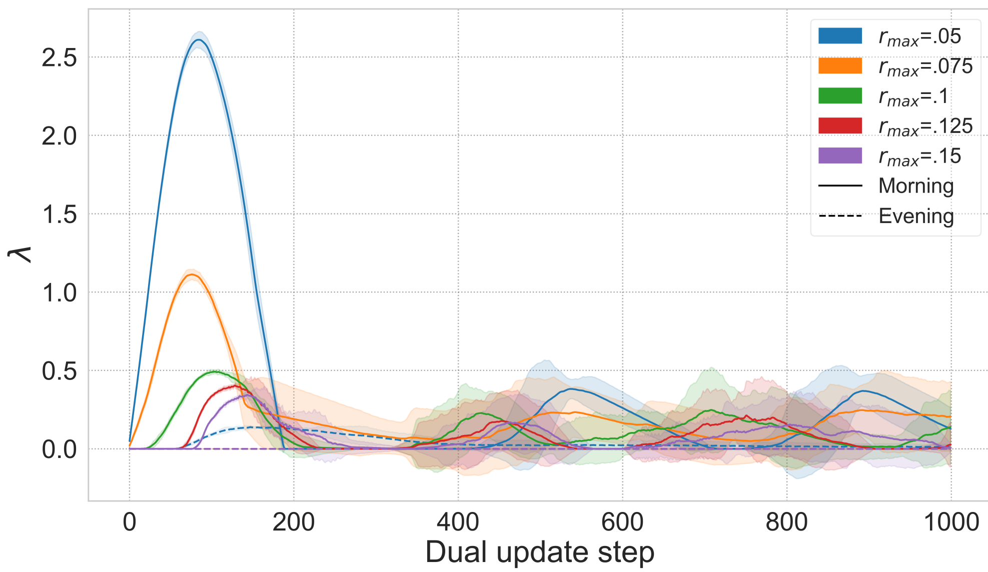

Figure S2: Dual variable λ for category 0 over training, for the five tightest bounds, morning (solid) and evening (dashed); bands show ± 1 standard deviation over 10 seeds

S3 Network coverage and service-zone table

The service zones are built in Section 5.1 of the main text and ordered into categories in Section 5.2. This section adds the supporting coverage figures and the full per-zone table.

S3.1 Network coverage

The 500 m radius used to weight stations sits past the elbow of the coverage curve in Figure S3, where the curve has flattened. Only 12% of addresses have no station within 500 m, and at that range most have five or more, which raises a modeling question: whether to attribute an address's demand to its nearest station or spread it across the nearby ones, that a station-level model would have to settle. The zone-level model used here is less exposed, since stations in a zone share one inventory and the construction keeps addresses' stations inside their own zone (main text, Section 5.1).

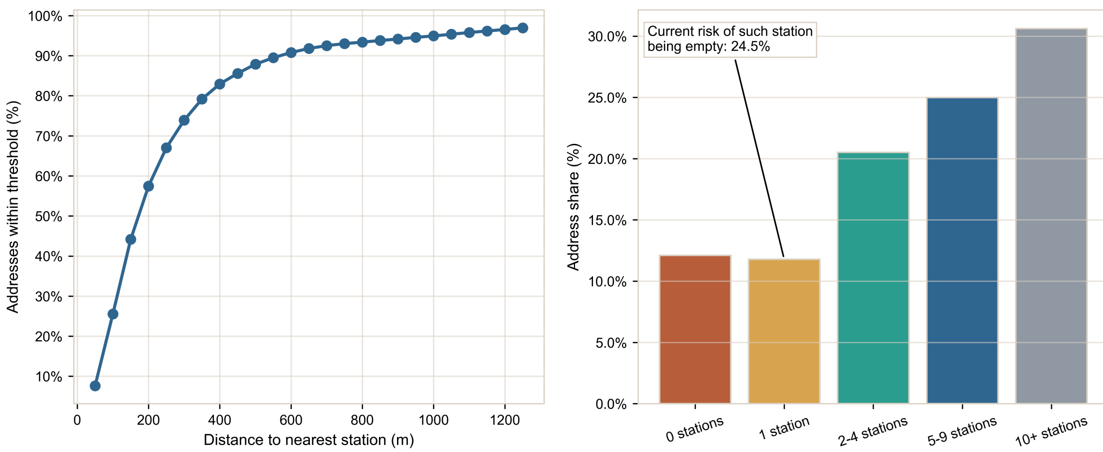

Figure S3: Left: The Hague addresses with a station within a given distance; Right: address share by number of stations within 500 m

The network has grown steadily between 2021 and 2026, with available bikes, address coverage, and the average distance from an address to its nearest bike all improving (Figure S4). These are descriptive checks; realized shared-bike trips were not available, so they motivate rather than validate the design.

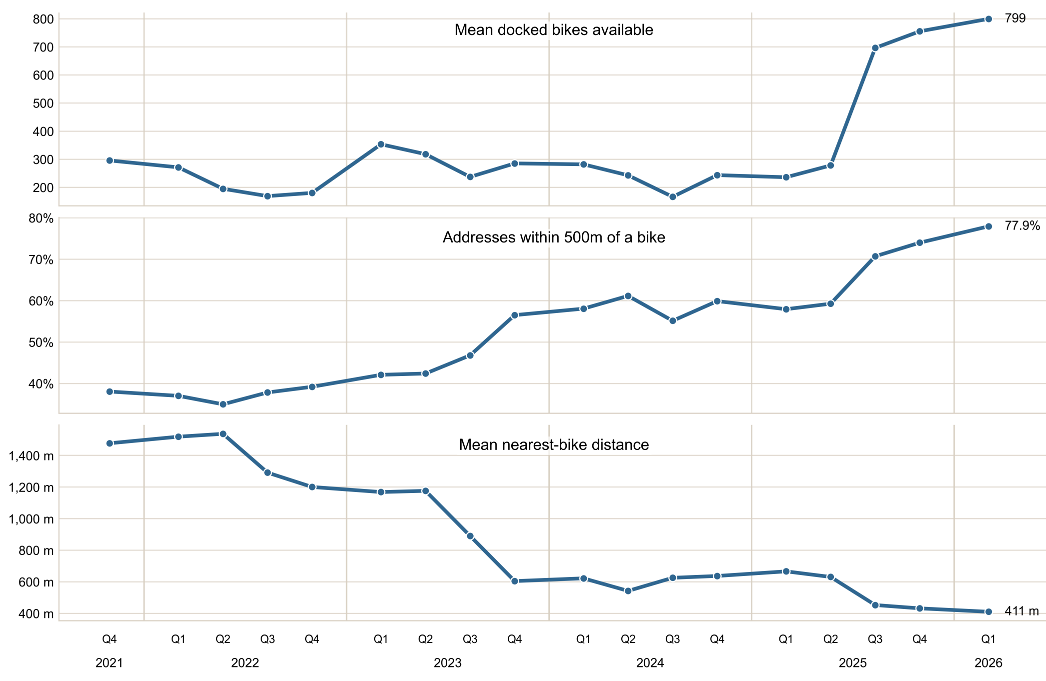

Figure S4: The Hague docked Donkey Republic bike-share access, 2021–2026

Table S1: Service zones by category, with station count, aggregate capacity, the three normalized measures ($\tilde{\lambda}_d$ departures, $\tilde{\rho}$ address density, $\tilde{\beta}$ non-residential activity), and the score

Zone	Category	Stations	Capacity	$\tilde{\lambda}_d$	$\tilde{\rho}$	$\tilde{\beta}$	Score
10	0	4	12	0.166	0.000	0.007	0.057
5	0	24	115	0.000	0.278	0.000	0.093
0	0	6	38	0.228	0.080	0.013	0.107
17	0	18	75	0.129	0.179	0.024	0.110
15	1	7	41	0.348	0.117	0.041	0.169
16	1	24	193	0.379	0.143	0.127	0.216
18	1	30	223	0.353	0.188	0.153	0.231
7	1	11	65	0.154	0.432	0.149	0.245
14	2	21	81	0.585	0.138	0.013	0.245
6	2	32	126	0.209	0.417	0.141	0.256
9	2	7	28	0.636	0.120	0.031	0.262
8	2	18	180	0.419	0.419	0.276	0.372
13	3	46	175	0.380	0.643	0.193	0.405
11	3	35	213	0.444	0.303	0.530	0.426
2	3	17	83	0.919	0.380	0.092	0.464
19	3	21	92	0.382	0.718	0.301	0.467
3	4	50	219	0.281	1.000	0.544	0.608
1	4	28	100	0.919	0.577	0.360	0.619
4	4	36	165	0.655	0.753	0.669	0.692
12	4	61	572	1.000	0.899	1.000	0.966

S3.2 Category assignment

Table S1 lists the twenty zones with their category, station count, aggregate capacity, the three normalized measures, and the score. For each zone z the score averages three measures, each normalized across the twenty zones,

$$\text{score}_z = \tfrac{1}{3}\big(\tilde{\lambda}_{d,z} + \tilde{\rho}_z + \tilde{\beta}_z\big), \tag{S1}$$

with $\tilde{\lambda}_{d,z}$ the normalized hourly departure rate from the Dutch national travel survey (ODiN), $\tilde{\rho}_z$ the normalized residential address density, and $\tilde{\beta}_z$ the normalized count of non-residential activities. The zones are sorted by their score from lowest to highest and split into five groups of four, the four lowest-scoring forming category 0 and the four highest category 4. Reproducing the skewed station counts of the synthetic scenario (60:40:30:20:10 across categories 0–4) while keeping enough zones in the smallest category would need far more than twenty zones, spreading the survey's trips too thinly across them to estimate demand reliably. The capacity and category are what the model uses. The station count and the three measures are listed for context.

S4 Demand and reward adaptations

This section covers the model adaptations for the case study: the demand process, the per-zone transition, and the reward and training changes.

S4.1 Demand realization

The baseline of Cederle et al. [1] approximates the hourly inventory transition with a Skellam net flow: each hour draws a single net change (arrivals minus departures) from a Skellam distribution, the difference of two independent Poisson variables. A net flow suffices when only the balance matters, but a single net value cannot say how many arrivals and departures produced it: an hour with two arrivals and none leaving looks identical to one with twelve arrivals and ten departures. This matters for failure counting, since in any hour that holds both arrivals and departures a departure that finds the zone empty fails even if an arrival follows.

The case study therefore models arrivals and departures as two independent Poisson processes. For each zone i and clock-hour h in period p,

$$A_{i,h} \sim \text{Poisson}\left(\lambda_{i,a}^{(p)}\right), \qquad D_{i,h} \sim \text{Poisson}\left(\lambda_{i,d}^{(p)}\right). \tag{S2}$$

The $A_{i,h}$ arrivals and $D_{i,h}$ departures are put into one list, shuffled uniformly at random, and applied in turn to the running inventory v_i: an arrival sets $v_i \leftarrow \min(C_i,\ v_i + 1)$; a served departure sets $v_i \leftarrow v_i - 1$; a departure at $v_i = 0$ leaves the inventory at zero and increments the failure count f_i. The end-of-period inventory has no closed-form distribution and follows from these rules.

S4.2 Per-zone transition

The dynamics combine a deterministic rebalancing step with the stochastic demand of Section S4.1. At a rebalancing instant the chosen occupancy action a_i of the main text is converted to vehicles, $\Delta v_i = \text{round}(a_i C_i)$, and applied with clipping,

$$v_i^+ = \min\big(C_i,\ \max(0,\ v_i + \Delta v_i)\big), \tag{S3}$$

where $v_i = \text{round}(o_i C_i)$ is recovered from the binned occupancy. The twelve hours of the period are then simulated event by event from v_i^+ as in Section S4.1, yielding the end-of-period count v_i' and the failure count f_i. The next state is

$$o_i' = \text{round}_{0.01}\left(\frac{v_i'}{C_i}\right), \qquad p' = 1 - p, \tag{S4}$$

so the period alternates between morning and evening. The transition kernel has no closed form and is defined by these steps.

S4.3 Reward rescaling

The reward terms of the main-text formulation are written for the synthetic capacity of 100 and are rescaled to occupancy units. The rebalancing cost is charged on the realized vehicle change:

$$\text{reb}_i = \alpha\, \phi(\kappa_i)\, \mathbf{1}[\Delta v_i \neq 0], \tag{S5}$$

which agrees with the indicator $\mathbf{1}[a_i \neq 0]$ except when rounding gives $\Delta v_i = 0$ for a nonzero action, possible in small zones. The fleet penalty keeps the category- and period-specific target $\mu_{\kappa,p}$ and tolerance $\zeta_{\kappa,p}$ of Cederle et al. [1], reinterpreted as percentages of capacity. As in the synthetic case, it is evaluated at the post-action occupancy $o_i^+ = v_i^+/C_i$, with v_i^+ from Section S4.2,

$$\ell_i(o_i^+, p) = \max\left(0,\ |o_i^+ - \frac{\mu_{\kappa_i,p}}{100}| - \frac{\zeta_{\kappa_i,p}}{100}\right) \cdot 100, \tag{S6}$$

where the factor of 100 restores the bike-unit scale of the synthetic case, so that α and ξ stay comparable across the two settings.

S4.4 Exploration-rate rescaling

The per-category ε-greedy decay is rescaled for the four-zone-per-category structure of the case study, well below the synthetic range of roughly 10 to 60 zones per category. A category with fewer zones produces fewer Q-updates per simulated day, so a fixed per-update decay would leave exploration too high at the end of training. The case study sets

$$\varepsilon_{\text{decay}}^{(\kappa)} = \varepsilon_{\text{decay}}^{0} \frac{n_\kappa^{\text{ref}}}{|\mathcal{V}_\kappa|}, \tag{S7}$$

with $\varepsilon_{\text{decay}}^0 = 8.25 \times 10^{-7}$ (the main-text value) and n_κ^{ref} the synthetic station count of category κ, so the total decay per simulated day matches the corresponding synthetic category. All other training parameters (n_{dual}, η_λ, γ, α, ξ) are unchanged.

References

[1] Cederle, M., Piron, L.V., Ceccon, M., Chiariotti, F., Fabris, A., Fabris, M., Susto, G.A.: A fairness-oriented reinforcement learning approach for the operation and control of shared micromobility services. In: 2025 American Control Conference (ACC), pp. 565–570. IEEE, Denver, CO, USA (2025). https://doi.org/10.23919/ACC63710.2025.11107820

[2] Stooke, A., Achiam, J., Abbeel, P.: Responsive safety in reinforcement learning by PID lagrangian methods. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20. JMLR.org, Brookline, MA, USA (2020)