City Builder: Does Induced Demand Make Transit Network Design a Reinforcement Learning Problem?

Daniel Mottesi | Stanford CS 224R | *Anticipation vs. Adaptivity Under Endogenous Demand*

Extended Abstract

Motivation. Transit design is almost always solved as a *static* optimization: fix today's demand, then choose the routes that serve it best. But infrastructure *changes* demand — well-served corridors attract population and jobs (induced demand / land-use feedback). Today's network reshapes tomorrow's city, making design a sequential decision under *endogenous, policy-dependent* demand: build now while anticipating how the city will grow in reaction. This poses a sharp question that reinforcement learning is positioned to answer: *how strong must the induced-demand feedback be — and what must the planner commit to — before anticipating it (RL) beats reacting to it (greedy)?* Induced demand is exactly what a myopic planner cannot see, so the setting is a clean test of anticipation against adaptivity.

Method. I formulate transit design as a multi-year MDP. The state is a city graph (streets, per-zone activity, current routes, induced demand); each year the agent builds route(s); the transition grows zone activity in proportion to the *transit-derived* accessibility the agent creates, then regenerates demand through a gravity model. The reward is integrated welfare $-\sum_t C(s_t)$. A single scalar α sets feedback strength and recovers the static setting of Holliday et al. (2024) at α=0. A GNN policy trained with PPO is compared against an *adaptive greedy* planner (re-optimizes the true cost each year) and random, under two decision regimes: *rebuild-yearly* and *build-then-watch*.

Implementation. I fork Holliday et al.'s transit-RL code as a *library* — reusing its validated welfare evaluator and graph schema, replacing the static MDP, rollout, and training loop — and add the closed induced-demand loop and two-regime environment. Two fixes were load-bearing: a reward cost-sizing bug (training scored the *planned* route count, evaluation the *realized* count), caught by a train-vs-eval parity test; and a PPO instability from frozen feature-normalizers under non-stationary demand, fixed by unfreezing the running statistics and clamping. Training runs on a modern `torch 2.8 / PyG 2.7` stack on cloud GPUs.

Results. Under rebuild-yearly, RL trails adaptive greedy and the deficit *grows* with α, falling below random: when re-planning is free, adaptivity dominates and the rich-get-richer dynamics pile induced demand onto already-served zones. Under build-then-watch the result *flips*: among the seeds that train cleanly, as α grows the RL–greedy gap *closes* and the best-of-N policy reaches and, at α=1.0 with sufficient training, robustly *exceeds* greedy across three seeds. Under short training budgets the outcome is bimodal, where some seeds collapse before a welfare phase-transition, but longer training recovers them.

Discussion. Induced demand alone does not make RL beat planning — the *decision regime* decides. Anticipation pays only once the planner is committed and cannot re-optimize its mistakes away. The α=1.0 win is robust across three seeds and holds for the *deterministic* policy in two of three (so it is more than a best-of-N effect), though the average sampled rollout is only near parity; and training is brittle under short budgets, where some seeds collapse exactly when watch-phase demand grows — the regime where anticipation should matter most.

Conclusion. I contribute a closed-loop induced-demand environment that is exact at α=0, an α-ablation that measures when anticipation pays, and the regime-crossing finding that *commitment* , but not feedback strength alone, governs the answer. Next steps: a behavior-cloning warm-start from greedy, calibration of α to Bogotá's TransMilenio rollout, and a Bogotá Metro Line 1 counterfactual under the calibrated dynamics.

City Builder: Does Induced Demand Make Transit Network Design a Reinforcement Learning Problem?

Daniel Mottesi
Stanford University
Department of Computer Science
mottesid@stanford.edu

Abstract

Transit network design is conventionally posed as static optimization against a fixed demand matrix, yet transit infrastructure reshapes the demand it serves through induced demand and land-use response. I recast network design as a multi-year MDP with an endogenous, policy-dependent demand process and ask a single question: how strong must the induced-demand feedback be, and what must the planner commit to, before a learned anticipating policy (a GNN trained with PPO) beats an adaptive greedy planner? I build a closed-loop simulator in which the agent's routes drive zone growth through transit-derived accessibility, calibrated so that a feedback knob α recovers the static setting of Holliday et al. (2024) exactly at α=0. Sweeping α under two decision regimes yields the paper's contribution: when the planner can re-plan every year, adaptive greedy wins and its advantage *grows* with α; but when routes must be committed and the city then evolves around them ("build-then-watch"), the gap *closes* with α and, among the seeds that train cleanly, the best learned policy reaches and slightly exceeds greedy at strong feedback. The regimes cross. Induced demand alone is not sufficient to make planning a reinforcement-learning problem — commitment is. There are limits: the α=1.0 win is robust across three seeds — and holds for the deterministic policy in two of three, so it is not merely a best-of-N effect — though the sampled mean is only near parity; and at short training budgets a fraction of seeds collapse before a welfare phase-transition that longer training clears.

1 Introduction

The transit network design problem (TNDP), to choose a set of routes over a city to move people well at acceptable operating cost, is NP-hard and has a long history of metaheuristic and, recently, learning-based solutions Guihaire and Hao (2008); Holliday et al. (2024). Almost all of this work treats travel demand as exogenous and fixed: an origin–destination matrix is given, and the planner optimizes against it. This is a convenient fiction. Decades of empirical transportation research document the opposite: building capacity *induces* travel, and well-served corridors attract population and economic activity over the following years and decades Duranton and Turner (2011); Cervero (2002). The network you build today changes the demand you face tomorrow.

Stanford CS224R 2025 Final Report

Once demand is endogenous and depends on the planner's own past decisions, network design stops being a one-shot optimization and becomes a sequential decision problem under policy-dependent dynamics — exactly the structure reinforcement learning is built for. This reframing also sharpens into a clean scientific question. A myopic, greedy planner optimizes for the demand it currently observes; induced demand is precisely the part of the future it cannot see. If the feedback is strong, a planner that *anticipates* how its routes will reshape the city should beat one that only *reacts*. So:

> *How strong must the induced-demand feedback be — and what must the planner be forced to commit to — before anticipating it (RL) beats reacting to it (greedy)?*

I answer this with a controlled simulation study. I construct a multi-year MDP whose transition closes an induced-demand loop. The agent's transit network determines accessibility, which drives zone growth, which regenerates demand, and is governed by a single feedback-strength parameter α that reduces the environment to a fixed-demand static setting at α=0. I then ablate α, comparing a GNN policy trained with PPO against an adaptive greedy baseline, under two decision regimes that differ only in whether the planner is allowed to re-plan each year.

My contributions are:

1. **A closed-loop induced-demand environment for transit design** that wraps a published, numerically validated welfare oracle, exposes feedback strength as a single knob α, and recovers the static baseline byte-for-byte at α=0 (Sections 3, 4).
2. **An α-ablation** measuring the RL-versus-greedy welfare gap as a function of feedback strength, under two decision regimes (rebuild-yearly and build-then-watch).
3. **The regime-crossing finding**: induced demand alone does not make RL beat greedy — the *decision regime* does. When re-planning is free the gap widens with α in greedy's favor; when infrastructure must be committed, the gap closes with α and the best learned policy reaches greedy (Section 5).
4. **An engineering contribution**: a correctness gauntlet (a train-vs-eval reward parity test and a stabilization of PPO under non-stationary demand) without which the comparison is not even well-posed (Section 4).

The study uses a single benchmark instance with a handful of seeds, and the strongest claim is that anticipation reaches and beats greedy *when training converges*. I treat the results as a map of *where* anticipation begins to pay, not as a claim that RL is ready to design real transit networks.

2 Related Work

Transit network design. The TNDP and its many variants are reviewed comprehensively by Guihaire and Hao Guihaire and Hao (2008). Classical approaches use metaheuristics over hand-designed neighborhood operators, and are benchmarked on small synthetic cities such as Mandl's network Mandl (1980) and the larger Mumford instances Mumford (2013). These methods assume a fixed demand matrix.

Learning-based transit design. Holliday, El-Geneidy, and Dudek Holliday et al. (2024) train a graph-neural-network policy with deep RL to construct and improve transit networks, either standalone ("learned construction") or as low-level heuristics inside an evolutionary algorithm, achieving state-of-the-art results on the Mumford benchmark and on a real network in Laval, Canada (see also their neural-evolutionary variant Holliday and Dudek (2024)). Their setting, however, has *fixed* demand: there is nothing to anticipate, because the city does not respond to the network. I build directly on their work, reusing their welfare evaluator and city-graph representation as a library while replacing the single-shot, static MDP with my multi-year, endogenous-demand formulation (Section 3).

Induced demand and land use. That transport infrastructure induces travel and reshapes land use is well established empirically: Duranton and Turner's "fundamental law of road congestion" Duranton and Turner (2011) and Cervero's synthesis of induced-travel evidence Cervero (2002) are canonical, and accessibility-based land-use models trace back to Hansen Hansen (1959). The work closest to mine in spirit is Paulsen and Rich Paulsen and Rich (2024), who plan *sequential* bicycle-network expansions while explicitly accounting for induced demand over the investment horizon — but they solve a classical cost–benefit / mixed-integer optimization, not a learned policy, and so produce no transferable, lookahead-capable controller.

The gap. No prior work combines (a) a learned policy that can anticipate with (b) an endogenous-demand environment, and then measures how the answer changes with feedback strength and decision regime. That measurement is my contribution.

3 Method

3.1 The multi-year MDP

I model transit design over a horizon of T years as an MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$.

State. $s_t = (G_t, x_t, B_t)$, where $G_t = (V, E_{\text{street}}, E^t_{\text{built}})$ is the city graph over a fixed set of N zones (nodes) with a static street network and the transit edges built by year t; $x_t \in \mathbb{R}^N_+$ is per-zone activity (combined population and employment); and B_t is the remaining annual construction budget. The induced demand D_t is a deterministic function of x_t and travel times and is carried in the graph.

Action. Each year the agent builds one complete new transit route (a path of zones), materialized by Holliday's inner route-construction policy constrained to a one-route episode and subject to the year's capex budget; a *no-build* action is available. Building a route over several segment-level decisions (start / extend / halt) lets me reuse the pretrained GNN route-construction machinery unchanged.

Transition. Applying a_t updates the built-edge set and the travel-time matrix; then the land-use dynamics advance activity and the demand matrix is regenerated (Section 3.2). The only stochasticity is in optional land-use noise; the network update is deterministic given the action.

Reward. I optimize *integrated welfare* over the horizon,

$$R \;=\; -\sum_{t=0}^{T-1} C(s_t), \qquad C(s_t) \;=\; \sum_{ij} D^t_{ij}\,\tau^t_{ij} \;+\; \lambda_{\text{op}}\,c_{\text{op}}(s_t) \;+\; \lambda_{\text{cov}}\,c_{\text{cov}}(s_t), \tag{1}$$

where the welfare cost $C(\cdot)$ is computed by the reused oracle as demand-weighted travel time ($\sum_{ij} D_{ij}\tau_{ij}$) plus an operator route-time term c_{op} and a coverage/unserved-demand penalty c_{cov}, and τ^t are network travel times. I discuss below (Section 4) why this *level-sum* functional is the correct objective and why the more obvious per-step welfare *delta* is not.

Horizon. T=12 years for the rebuild-yearly experiments and T=40 years (10 build + 30 watch) for build-then-watch (Section 3.5).

3.2 Closing the induced-demand loop

Zone activity evolves multiplicatively in accessibility:

$$x_{t+1,i} \;=\; \text{clip}\Big(x_{t,i}\cdot\big(\text{base} + \alpha\,\tilde{A}_{t,i}\big) + \varepsilon_{t,i},\; 0,\; \text{cap}_i\Big), \tag{2}$$

where $\tilde{A}_{t,i} \in [0,1]$ is the min–max-normalized Hansen accessibility Hansen (1959) (β=2) of zone i under the current network, $\alpha \geq 0$ is the induced-demand strength, ε is optional

Gaussian noise, and base=1 so that α=0 **with** ε=0 **is the byte-exact identity** — the property that lets me reproduce the static baseline and use it as an integration-correctness gate. After growth, demand is regenerated by a gravity model $D_{ij} \propto x_i x_j \, d_{ij}^{-\beta}$ Wilson (1971).

Two design choices close the loop so that the agent's network genuinely shapes the city:

- **Growth keys off *transit* accessibility, not street accessibility.** Land-use growth is driven by transit OD times (route graph plus transfer penalties), while the gravity demand model stays on street times. On a uniform-speed model a constrained transit path is always at least as long as the direct street path, so $\min(\text{street}, \text{transit})$ collapses to street and gives the agent no lever; transit-only accessibility makes the network fully control who becomes accessible, so growth bends toward served corridors.
- **Latent demand stays network-independent.** Keeping the gravity model on street times means unserved pairs still carry demand, so failing to serve them is penalized — preserving the incentive to expand.

A subtlety I return to in the discussion: under the default dynamics the capacity cap is $\text{cap}_i = 3x_{0,i}$, tied to a zone's initial size. This makes the process *rich-get-richer* — already-large zones can grow most — which is exactly why anticipation struggles in the rebuild-yearly regime. I therefore expose two objective-preserving knobs, `cap_blend` (decouple the cap from initial size) and `add_rate` (size-independent growth), so that connecting an underserved zone can *create* demand a greedy planner would not foresee.

3.3 Policy and training

The policy is Holliday's GNN encoder, which reads the city graph and builds each route segment-by-segment (start / extend / halt). Demand enters the network through edge features that are recomputed at every step from the live graph, so mutating demand between years propagates to the policy with no stale-feature problem. I train with PPO Schulman et al. (2017): clipped objective (ϵ=0.2), generalized advantage estimation Schulman et al. (2016) (λ=0.95, γ=0.95), advantages normalized per pooled buffer, two epochs per update, minibatch 32, two episodes pooled per update, entropy coefficient 0.01, Adam (lr 3×10^{-4}, decay 8.4×10^{-4}) with a neural value baseline (lr 5×10^{-4}), and rewards/returns scaled by 0.01 for critic conditioning. I run 150 PPO iterations with best-model selection on integrated welfare; checkpoints are persisted on every improvement so cloud preemption never loses the best policy.

3.4 Baselines

The central comparison is against an **adaptive greedy** planner that, each year, re-optimizes the true welfare cost and adds the best marginal route — a strong baseline, because it sees the realized demand each year and reacts to it. I also report a **random** route baseline as the lower anchor. RL routes are replayed through the identical environment path used to score the baselines, so the comparison is fair.

3.5 Two decision regimes

The same MDP is evaluated under two regimes that differ only in *when the planner may act*:

- **Rebuild-yearly:** the agent (and greedy) re-plan every year over a 12-year horizon. Re-planning is free, so mistakes can be corrected as demand materializes.
- **Build-then-watch:** routes are committed in years 0–9, the network is then *frozen*, and the city evolves for 30 more years (40-year horizon) while welfare is integrated over the full horizon. Early decisions are judged by demand that materializes *after* the planner can act — the realistic case for metro/BRT infrastructure that lasts decades. Baselines obey the same build window.

Build-then-watch raises a credit-assignment problem: how does 30 years of frozen-network welfare get attributed to the 10 build-phase actions? I compare crediting schemes (a terminal

lump versus spreading the watch-phase reward across the build steps) and find the choice interacts with stability (Section 6).

4 Experimental Setup

From Mandl to Mumford0. I validated the environment before any RL. At α=0 the year-over-year welfare drift is 0.000, confirming the integration is correct and the static baseline is recovered. I found the small Mandl network Mandl (1980) unsuitable as a testbed: its welfare cost *floors after roughly two routes*, so the network saturates demand before induced growth can act, leaving no headroom for anticipation to matter. I therefore moved all headline experiments to **Mumford0** Mumford (2013) (30 nodes, a 12-route benchmark with longer routes), which saturates more slowly so that induced demand can reshape the city over the horizon. The knob behaves as designed: activity grows from 1.0× at α=0 (static) to ∼2.7× at α=2, and the fraction of zones saturated by year 10 rises from 0.07 to 0.80; I adopt α=0.5 as a reference operating point and sweep a principled range.

Protocol and metric. I run three seeds per α for the headline runs (converged build-then-watch and rebuild-yearly); an earlier short-budget build-then-watch sweep used four. Following Holliday's LC-100 evaluation, each policy is sampled 100 times and I report both the *best-of-N* rollout (the headline, since route construction is stochastic) and the *mean-of-N* with its seed spread. I also report the deterministic argmax decode as a convergence diagnostic. To compare across α despite the welfare scale changing with city size, I use a **scale-normalized gap**

$$\text{gap} \;=\; \frac{\text{RL} - \text{greedy}}{\text{greedy} - \text{random}}, \tag{3}$$

so that 0 means parity with greedy and −1 means random-level; positive means beating greedy.

The correctness gauntlet. Before any RL-vs-greedy number is meaningful I hardened the pipeline. *(i) A reward cost-sizing bug.* Training computed cost on the live planning state, whose "number of routes to plan" equals the full horizon T, while evaluation sized the cost to the *realized* route count; the empty planned-route slots inflated the training cost, so the policy optimized a different, non-stationary surface than greedy is scored against. A train-vs-eval parity test (driving one fixed route sequence through both paths and asserting equal per-year cost) now guards against this. *(ii) The reward functional itself.* The natural per-step welfare *delta* reward telescopes to $C(s_0) - C(s_T)$, which I found is nearly α-insensitive and, perversely, ranks random above greedy; the integrated level-sum of Eq. 1 is α-sensitive and ranks greedy above random, so it is the objective I adopt. *(iii) PPO instability.* Holliday freezes the GNN's feature-normalizers on iteration-0 statistics, which is correct for static demand but wrong here: a trained policy grows demand several-fold, so the frozen normalizer under-normalizes and the GNN's node-pair scores overflow (worse at high α). Unfreezing the running statistics and clamping the scores fixes it. *(iv) A decode audit* confirmed the argmax "collapse" below the sampled mean is not a bug but the expected symptom of a diffuse, under-converged policy.

Stack and compute. I rebuilt Holliday's environment (originally `torch 1.12 / PyG 2.2 / Python 3.9`) on `torch 2.8 / PyG 2.7 / Python 3.12` and reproduced the published Mandl LC-100 numbers within seed variance, the credibility anchor for everything downstream. RL training runs on A10G GPUs via Modal.

5 Results

5.1 Result 1 — Rebuild-yearly: the gap widens with feedback

When the planner re-plans every year, my hypothesis was that the gap would open in RL's favor as induced demand strengthens; the opposite happens (Table 1). RL sits between greedy and random at α=0 and degrades with α, with even best-of-N dropping below random

Table 1: Rebuild-yearly RL−greedy gap vs. α (Mumford0, n=3). RL trails the adaptive greedy baseline and the deficit grows with α.

α	norm. gap (best)	norm. gap (mean ± sd)	raw gap (mean)
0	−0.26	-0.59 ± 0.18	−7.0
0.5	−0.82	-1.43 ± 0.80	−20.8
1.0	−1.36	-1.75 ± 1.19	−25.5

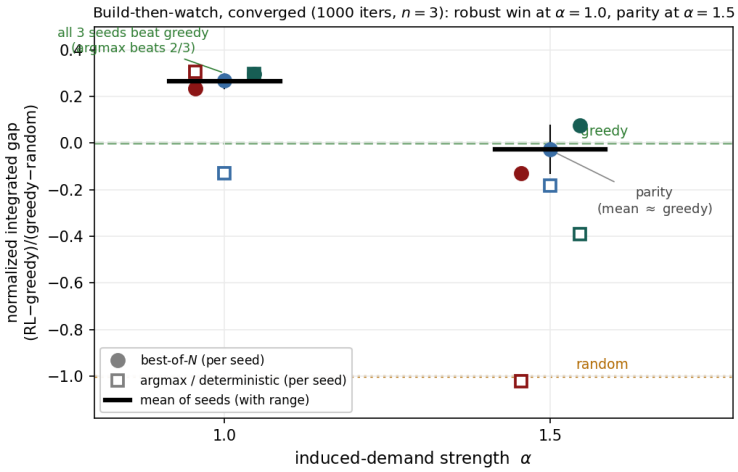

Figure 1: **Converged build-then-watch** (1000 iterations, n=3; normalized integrated gap, 0=greedy, −1=random). At α=1.0 all three seeds beat greedy in best-of-N (filled) and the deterministic argmax (open) beats greedy in two of three; at α=1.5 the seeds straddle greedy (parity). Black bar = seed mean with range.

by α=1.0 — a real α-trend, not a welfare-scaling artifact (the normalized metric controls for scale). Two causes compound: *adaptivity dominates when re-planning is free* (greedy re-optimizes every year, absorbing induced growth as it appears), and *the dynamics are rich-get-richer* (with a cap tied to initial size, induced demand piles onto already-served zones while peripheral zones cannot grow, so there is no "future demand greedy ignored" to anticipate). This motivated the second regime.

5.2 Result 2 — Build-then-watch: commitment flips the result

Committing routes early and freezing the network while the city evolves inverts the picture: the build-then-watch gap *closes* as α grows while the rebuild-yearly gap widens, so the two regimes **cross** — the paper's central finding. At strong feedback (α=1.0), with a sufficient training budget (1000 iterations, three seeds), the win is robust: all three seeds beat greedy in best-of-N (normalized integrated gap $+0.24, +0.27, +0.30$; mean $+0.27$), and the *deterministic* argmax policy beats greedy in two of three (mean $+0.16$) — so this is not merely a best-of-N tail effect; the converged policy itself is better. The average sampled rollout sits just below greedy (mean-of-N -0.07, $t = -4.2$): the headline is best-of-N, with the expected rollout near parity. Figure 1 also shows the strongest-feedback point I ran: at α=1.5 the advantage erodes to parity (mean -0.03, straddling greedy), partly because greedy itself improves as demand grows — so the build-then-watch advantage peaks at moderate-strong feedback rather than growing without bound.

Brittleness, and why the win needs enough training. The win depends on training budget. At a short 150-iteration budget some seeds collapsed to a single deterministic, *sub-random* network (sampling entropy $\to 0$), worst at high α and absent at α=0 — evidence that the non-stationary demand itself destabilizes training. This proved largely *under-training*: a

welfare phase-transition occurs several hundred steps in (around step 350–450), beyond the short budget, which the 1000-iteration runs clear. Making convergence dependable across seeds — an entropy floor, or a behavior-cloning warm-start from greedy — is the main remaining gap.

6 Discussion

The decision regime, not feedback strength, decides. The clean takeaway is that induced demand *alone* does not make transit design a reinforcement learning problem. When re-planning is free, an adaptive greedy planner that simply reacts each year is hard to beat, and stronger feedback only helps the reactor (it has more signal to react to). Only when the planner is *committed* — routes frozen while the city evolves — does anticipation have something to buy, and there the value of anticipation grows with feedback strength. The contribution is this flip, and it reframes the motivating question: the right question is not "is induced demand strong?" but "can the planner take its decisions back?".

Reliability is the open problem. The win is more than a best-of-N effect — at α=1.0 the converged policy's *deterministic* argmax beats greedy in two of three seeds — but it is not yet dependable. Under short budgets, non-stationary demand (coupled to the policy's own improving behavior) destabilizes training and some seeds collapse to a sub-random network; this is largely under-training (a longer fresh run clears the phase-transition and recovers the win), though resuming an *already*-collapsed checkpoint does not help. The levers for dependable convergence are an entropy floor and a behavior-cloning warm-start from greedy (annealed off, separating route *construction* from *anticipation*); the decode dispatch itself was audited and is correct.

Threats to validity. This is a single benchmark instance with few seeds, so the results are directional. The headline α=1.0 win rests on three converged seeds (all positive in best-of-N, two of three in argmax); its robustness to more seeds, more instances, and other feedback levels is untested, and at short training budgets convergence is unreliable. The rich-get-richer dynamics are a modeling choice (I expose knobs to relax them; the headline runs use `cap_blend`=`add_rate`= 0.5). I present the regime-crossing as a robust qualitative phenomenon and the α=1.0 magnitude as a three-seed point estimate from the converged (1000-iteration) run.

7 Conclusion

I recast transit network design as a multi-year MDP with endogenous, policy-dependent demand, built a closed induced-demand loop that recovers the static baseline exactly at α=0, and used it to ask when anticipation beats reaction. The answer is that the *decision regime* decides: under free re-planning the adaptive greedy baseline wins and its margin grows with feedback strength, but under commitment the gap closes with feedback strength: with sufficient training the learned policy robustly beats greedy at α=1.0 across three seeds (deterministically in two of three). Induced demand is necessary but not sufficient to make planning an RL problem — *commitment* is the missing ingredient. The immediate next steps target reliability: a sufficient training budget (a longer run already recovers the collapsed seeds into a robust three-seed win at α=1.0) plus an entropy floor and a behavior-cloning warm-start from greedy to make convergence dependable across seeds; then calibration of α to Bogotá's TransMilenio rollout and a Bogotá Metro Line 1 counterfactual under the calibrated dynamics (the simulator is under construction).

8 Team Contributions

This was a solo project.

References

Robert Cervero. 2002. Induced Travel Demand: Research Design, Empirical Evidence, and Normative Policies. *Journal of Planning Literature* 17, 1 (2002), 3–20.

Gilles Duranton and Matthew A. Turner. 2011. The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review* 101, 6 (2011), 2616–2652.

Valérie Guihaire and Jin-Kao Hao. 2008. Transit network design and scheduling: A global review. *Transportation Research Part A: Policy and Practice* 42, 10 (2008), 1251–1273.

Walter G. Hansen. 1959. How Accessibility Shapes Land Use. *Journal of the American Institute of Planners* 25, 2 (1959), 73–76.

Andrew Holliday and Gregory Dudek. 2024. A Neural-Evolutionary Algorithm for Autonomous Transit Network Design. *arXiv preprint arXiv:2403.07917* (2024). `https://arxiv.org/abs/2403.07917`

Andrew Holliday, Ahmed El-Geneidy, and Gregory Dudek. 2024. Learning Heuristics for Transit Network Design and Improvement with Deep Reinforcement Learning. *arXiv preprint arXiv:2404.05894* (2024). `https://arxiv.org/abs/2404.05894`

Christoph E. Mandl. 1980. Evaluation and optimization of urban public transportation networks. *European Journal of Operational Research* 5, 6 (1980), 396–404.

Christine L. Mumford. 2013. New heuristic and evolutionary operators for the multi-objective urban transit routing problem. In *IEEE Congress on Evolutionary Computation (CEC)*. 939–946.

Mads Paulsen and Jeppe Rich. 2024. Societally optimal expansion of bicycle networks. *Transportation Research Part B: Methodological* (2024). Sequential network expansion with induced demand, solved by cost–benefit optimization (non-RL). TODO: confirm volume/pages/year..

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations (ICLR)*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Alan G. Wilson. 1971. A family of spatial interaction models, and associated developments. *Environment and Planning A* 3, 1 (1971), 1–32.

A Implementation and Reproduction

The project forks Holliday's `transit_learning` repository; new code lives under `learning/city_builder/` (`land_use_dynamics.py`, `multi_year_mdp.py` with the `CityBuilderEnv`, `train_rl.py`, `eval_rl.py`, and `parity_train_vs_eval.py`), reusing the upstream welfare oracle (`MyCostModule`) and the `CityGraphData` schema. The α-ablation driver trains one policy per (α, seed) and evaluates each against greedy/random. The headline build-then-watch runs use `horizon_years=40`, `build_years=10`, `add_rate=0.5`, `cap_blend=0.5`, `reward_mode=integrated`, and `watch_credit=spread`; the converged result is at 1000 iterations. Full PPO hyperparameters are in Section 3. See the README in the github for more information.