RL Multi-Objetivo Aplicado ao Problema Integrado de Patrulha e Despacho

Moacir Almeida Simões Júnior
Tobias de Abreu Kuse

O Problema Central: Um Conflito de Objetivos

- Operações policiais têm dois objetivos conflitantes e simultâneos.
- **1. Patrulhamento Proativo:** Maximizar a *presença* em áreas de alto risco ("hotspots") para dissuadir o crime.
- **2. Despacho Reativo:** Minimizar o *tempo de resposta* a novas chamadas de emergência dinâmicas.
- **O Desafio:** Cada vez que uma unidade é despachada *reativamente*, ela compromete o plano de patrulha *proativo*.
- Este é um problema **bi-objetivo**.

Um Panorama das Soluções

- Problema complexo: mistura previsão espaço-temporal, roteamento e eventos estocásticos em tempo real.
- Três filosofias de solução principais:
  1. Modelos **Híbridos Heurísticos** (Simulação + Meta-heurísticas)
  2. Modelos **MARL Descentralizados** (RL Multiagente)
  3. Modelos **MORL Centralizados** (RL Multi-Objetivo)

Abordagem 1: Híbrida-Heurística

(Simões Júnior & Borenstein, 2025)

- **Conceito:** Modelar o sistema com simulação de eventos discretos.
- **Método:**
  - **Prever:** Usar ML (ex: XGBoost) para prever *hotspots dinâmicos*.
  - **Patrulhar:** Usar **Otimização por Colônia de Formigas (ACO)** para rotas de patrulha ideais.
  - **Testar:** Simular interrupções por chamadas aleatórias.
- **Conclusão:** Abordagem de *"otimizar e depois simular"*.

Abordagem 1: Método e Hotspots Dinâmicos

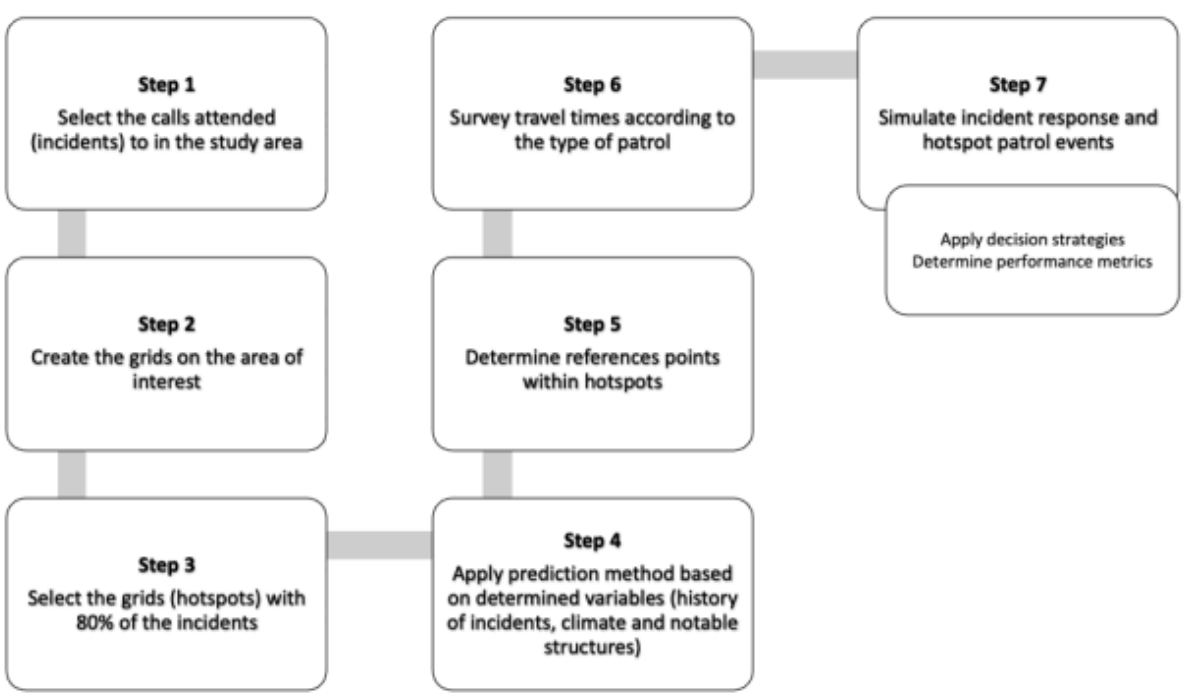


Figure 1: Method description

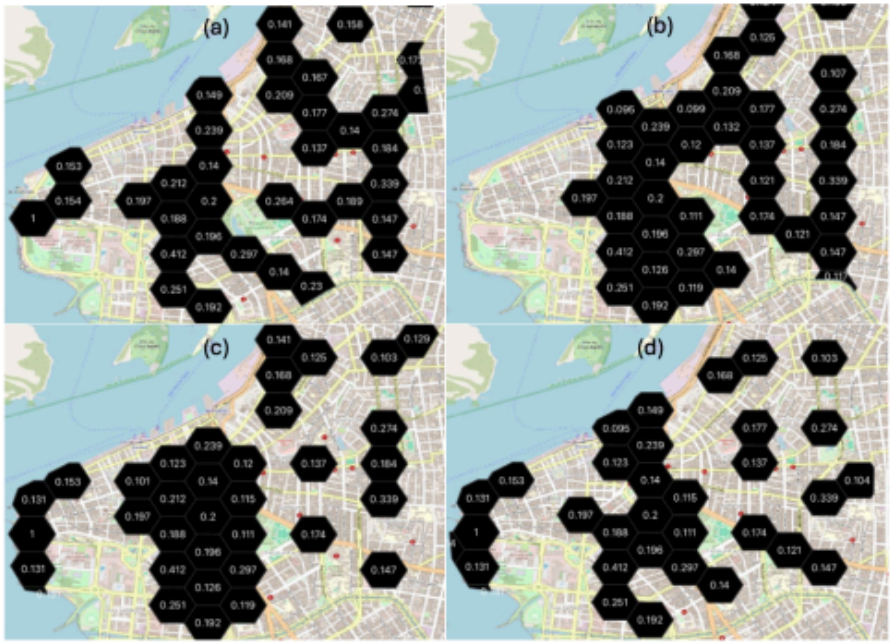


Figure 2: Evolution of crime hotspots throughout January 1, 2022, Porto Alegre downtown, Brazil. Subfigures (a), (b), (c), and (d) represent snapshots taken at 12:00 AM, 2:00 AM, 4:00 AM, and 6:00 AM, respectively. The figure highlights the top 50 hotspots ranked by their risk probabilities, with each hotspot labeled by its associated weight from the W set. This visualization illustrates how the predicted risk evolves dynamically, guiding police patrol actions.

Abordagem 2: Modelo Multiagente

(Repasky et al., 2024)

- **Conceito:** Cada viatura = um agente de RL.
- **Método:**
  - Sistema **heterogêneo** com $N + 1$ agentes.
  - N **Patrulheiros:** aprendem políticas próprias (DQN compartilhada).
  - **1 Despachante:** aprende política central (MIP + VFA).
- **Conclusão:** Abordagem **descentralizada**, com agentes aprendendo individualmente.

Abordagem 2: Visualização do Modelo MARL

Figure 1: The joint task of patrol and dispatch involves decisions such as that demonstrated above. It is unclear whether it is best to move one patroller across beat boundaries to manage an incident, leaving its beat un-managed, or to require the patroller responsible for that beat to travel a long distance to the scene.

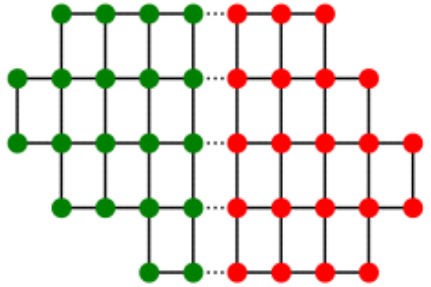


Figure 2: Example of a graph with two beats indicated by color. The dotted edges indicate that patrollers can travel between beats only when responding to calls.

Abordagem 3: RL Multi-Objetivo (MORL)

- RL padrão → recompensa escalar R.
- MORL → vetor de recompensas $\vec{R} = \langle R_{\text{resposta}}, R_{\text{presenç a}} \rangle$.
- **Desafio:** "Ótimo" é subjetivo (quanto vale piorar patrulha pra melhorar resposta?).
- **Solução comum:** Soma ponderada

$$R_{total} = w_1 R_{resposta} + w_2 R_{presenç a}$$

- **Problema:** Altamente sensível aos pesos — arbitrários e difíceis de justificar.

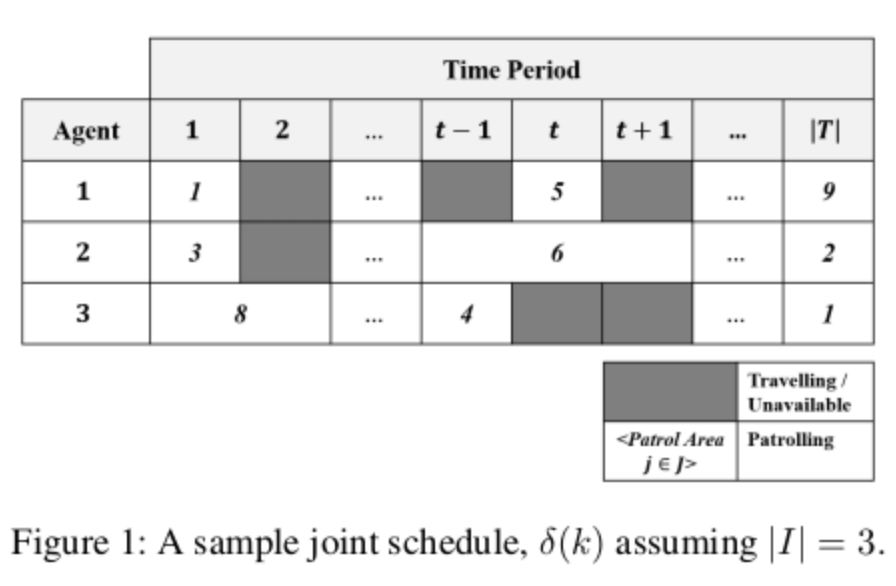
Visão Geral e Formulação

- Para contornar as dificuldades da abordagem de soma ponderada, o artigo de **Joe, Lau, & Pan (2022)** apresenta um modelo MORL centralizado que não utiliza essa técnica.
- **Título:** *Reinforcement Learning Approach to Solve Dynamic Bi-objective Police Patrol Dispatching and Rescheduling Problem*
- **Formulação:** MDP Centralizado de Agente Único.
- **Agente de RL:** Planejador central / despachante.
- **Recursos:** Unidades de patrulha (recursos controlados).
- Não é MARL — unidades não aprendem.

O MDP: Estado e Ação

- **Estado (S_k):** estado completo do sistema no incidente ω_k.
  - Hora atual (t_k), cronogramas ($\delta(k)$), status de patrulha ($\sigma(k)$).
- **Ação (x_k):** tupla de decisão $\langle x_k^i, x_k^t, \delta^x(k) \rangle$.
  - x_k^i: quem despachar
  - x_k^t: quando
  - $\delta^x(k)$: novo cronograma unificado
- **Insight:** espaço de ação enorme — decisão global sobre todas as unidades.

O MDP: Exemplo de Cronograma Conjunto ($\delta(k)$)



	Time Period							
Agent	**1**	**2**	...	$t-1$	t	$t+1$	...	$\|T\|$
1	*1*		...		*5*		...	*9*
2	*3*		...	*6*			...	*2*
3	*8*		...	*4*			...	*1*

	Travelling / Unavailable
<Patrol Area $j \in J$>	**Patrolling**

Figure 1: A sample joint schedule, $\delta(k)$ assuming $|I| = 3$.

A Solução: Híbrido de VFA + Heurística

- **Divisão de trabalho:**
  - **Offline (Módulo de Decisão):**
    - Rede de Função Valor $\hat{V}$ treinada offline.
    - Aprende o valor futuro esperado de cada estado.
  - **Online (Gerador de Ação):**
    - Heurística de Reagendamento (Ejection Chains) gera ações viáveis.
    - $\hat{V}$ avalia cada ação e escolhe:

$$x_k^* = \text{argmax}_{x_k}\{R_{imediata} + \gamma \hat{V}(S_k^x)\}$$

A Solução: Arquitetura do Modelo

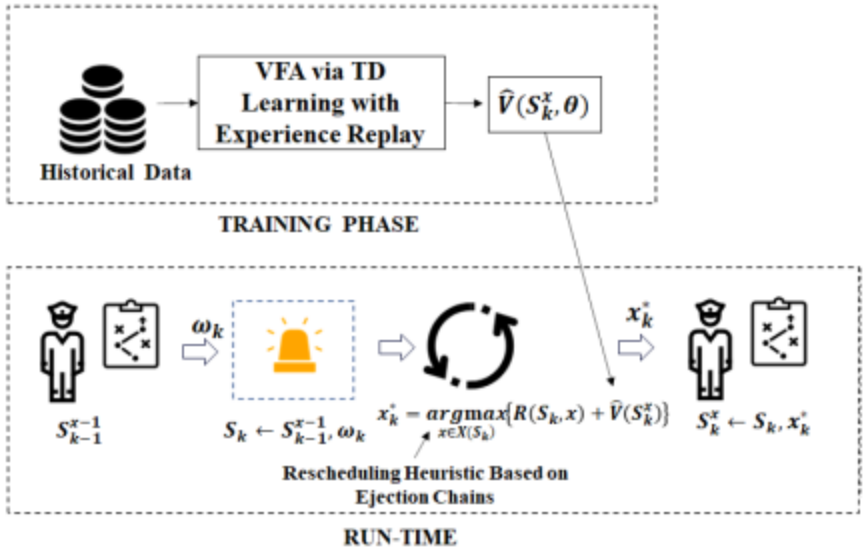


Figure 2: The learned value function approximation is utilised by the rescheduling heuristic to compute rescheduling decisions during run-time.

Inovação Bi-Objetivo (Sem Soma Ponderada)

- Recompensa **multiplicativa e diferencial**:

$$R(S_k, x_k) = f_r(x_k) \times f_p(\delta^x(k)) - f_p(\delta(k))$$

- **Componentes:**
  - $f_r(x_k)$ → sucesso da resposta (1.0, 0.5, 0)
  - $f_p(\delta^x(k)) - f_p(\delta(k))$ → mudança na presença da patrulha
- **Justificativa da Abordagem:**
  - A recompensa pondera *resposta* × *impacto na patrulha*.
  - Uma falha que destrói o plano de patrulha → fortemente penalizada.

Resultados Chave

- **Desempenho do Aprendizado:**
  - O modelo "integrado" (Joint) aprende mais rápido e atinge uma recompensa cumulativa maior e mais estável.
  - O método "Two-Stage" se mostra instável e inferior, especialmente em cenários complexos.
- **Taxa de Sucesso Final:**
  - O híbrido ($\hat{V}$ + heurística) supera estatisticamente o método "Two-Stage" na taxa de sucesso de resposta.

Resultados: Análise Gráfica

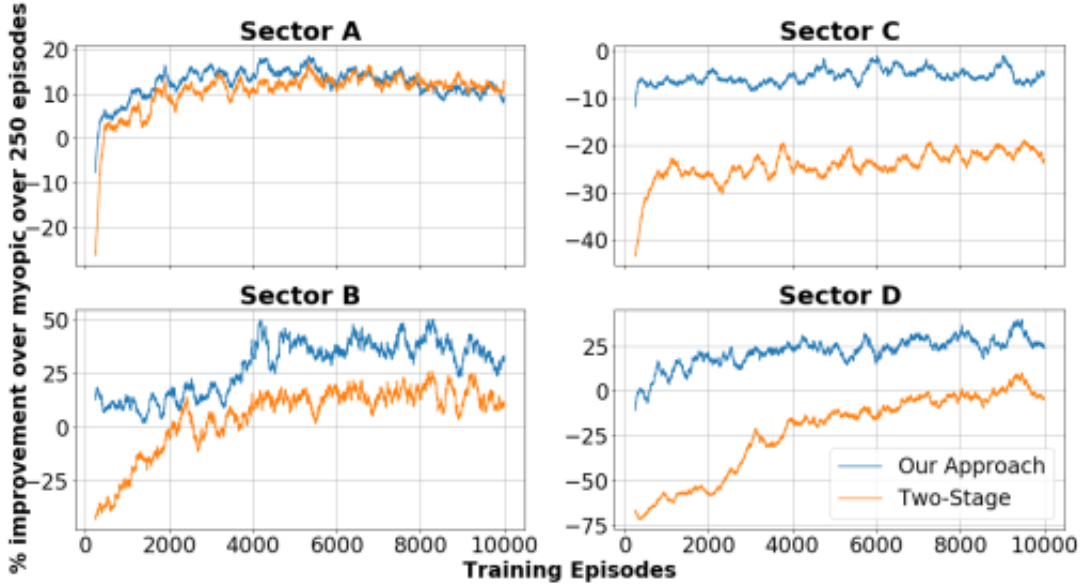


Figure 7: Average cumulative rewards over last 250 training episodes.

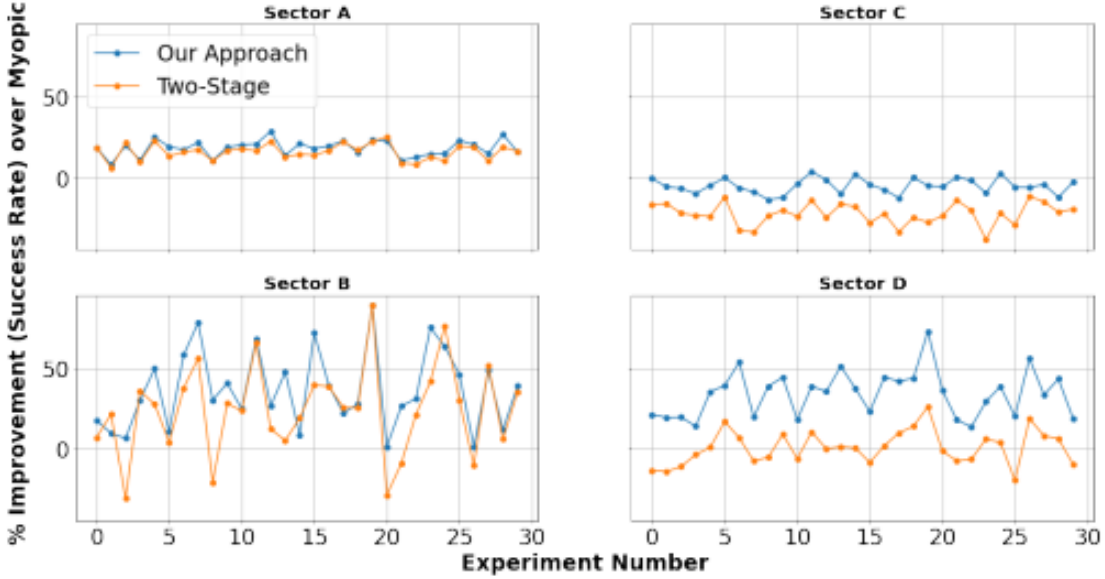


Figure 8: The gap in performance between our approach and the Two-Stage approach widens with more complex problem scenarios.

Resumo das Abordagens

- **Problema:** Patrulha Integrada (Proativa) & Despacho (Reativo).
- **Híbrida/Heurística:** poderosa, mas depende de heurísticas (ACO).
- **MARL:** baseado em agentes, mas decompõe o problema.
- **MORL Centralizado:** abordagem holística (VFA + heurística inteligente).

Principais Conclusões

- Problema real e ideal para RL avançado.
- Arquitetura de Joe, Lau & Pan:
  - **Gerador-Heurístico + Avaliador-VFA** → ótimo para espaços de ação complexos.
- Recompensa multiplicativa → resolve bi-objetivo sem pesos arbitrários.

Obrigado! 🙌

Perguntas?