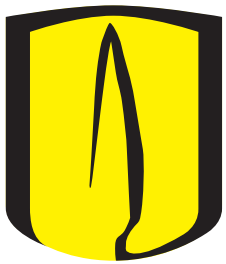
Control de intersecciones semaforizadas aplicando aprendizaje por refuerzo multiagente

Carolina Higuera Arias

Asesor: Ph.D Fernando Enrique Lozano Martinez

Universidad de los Andes

Diciembre 16, 2016



Motivación y definición del problema

Congestión vehicular en Bogotá

Causas	Consecuencias
• Incremento del parque automotor • Atraso en la infraestructura vial • Semáforos mal programados	• Altos tiempos de espera • Problemas económicos • Problemas ambientales

Pregunta de investigación

¿Cómo mejorar la infraestructura actual, de tal manera que se haga un uso inteligente y óptimo de los semáforos de la malla vial?

Solución propuesta

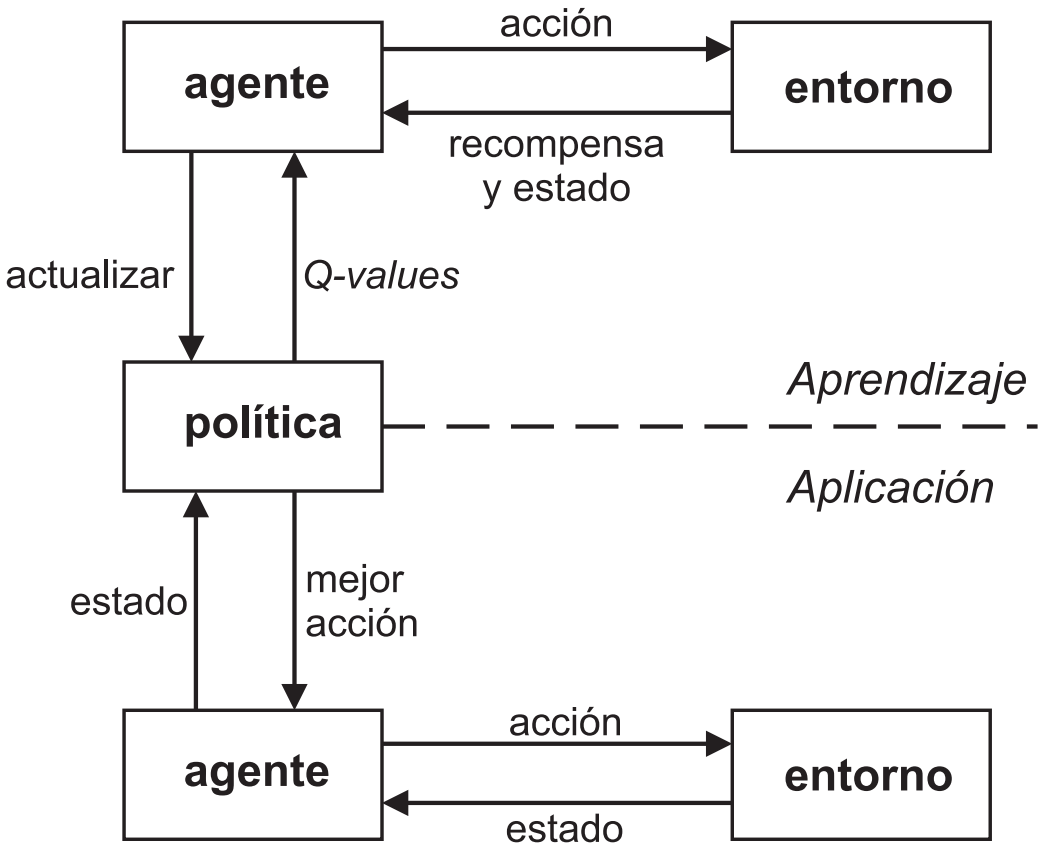
Obtener una estrategia de control con las siguientes características:

- Accionado por el tránsito
- Independiente del modelo matemático del sistema
- Que busque minimizar objetivos específicos para el sistema

} Aplicar aprendizaje por refuerzo multiagente (MARL)

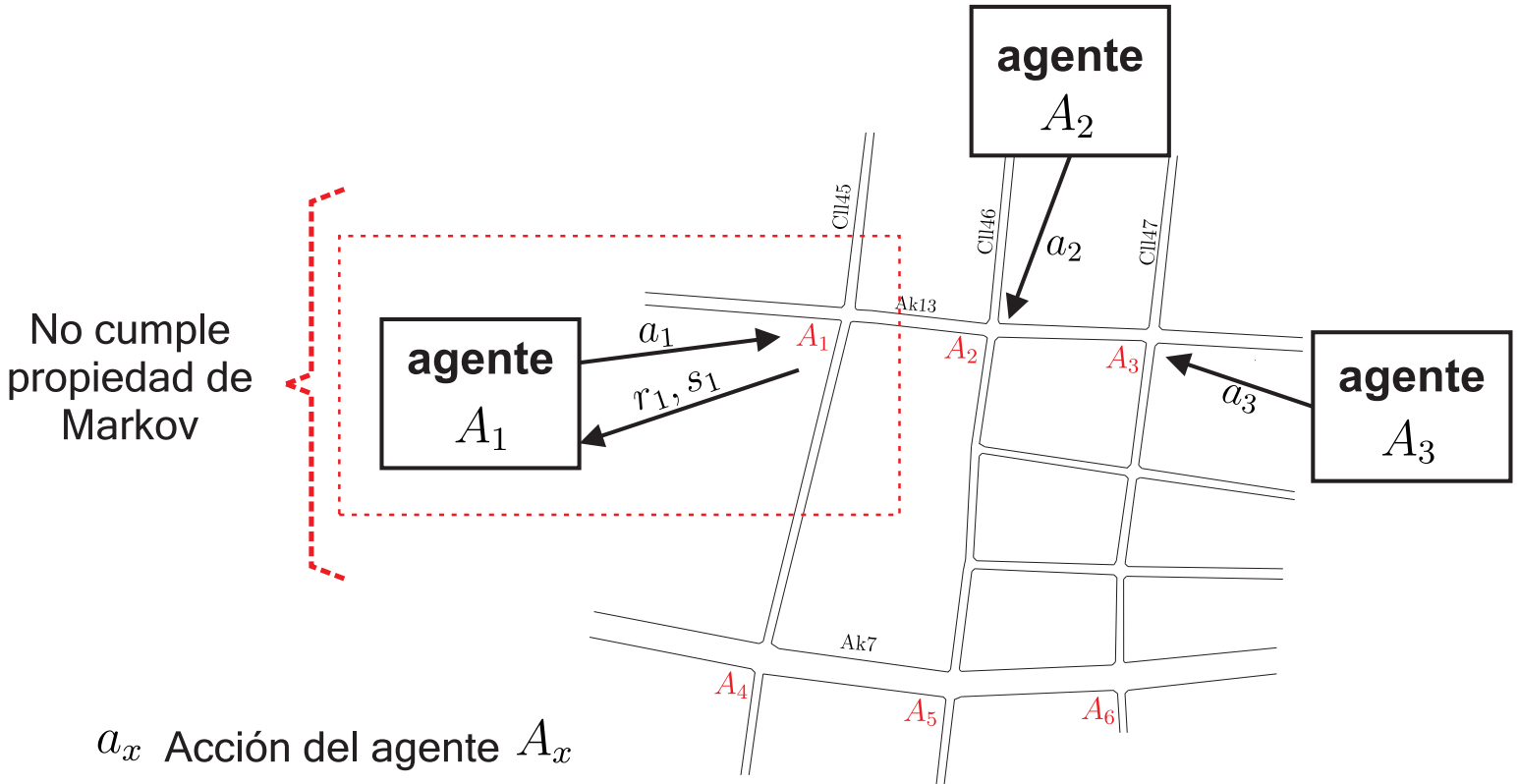
Aprendizaje por refuerzo

Caso de un solo agente:



Aprendizaje por refuerzo

Caso multiagente:



a_x Acción del agente A_x

r_i Recompensa recibida por el agente A_i

s_i Estado observado por el agente A_i

- Se puede describir todo el sistema como un MDP multiagente colaborativo.

Aprendizaje por refuerzo

Caso multiagente:

Sistema multiagente descrito principalmente por:

- Un tiempo discreto k
- Un grupo de n agentes $A_1, A_2, \cdots, A_n$
- Un conjunto finito de estados $\mathbf{s}^k \in \mathcal{S}$
- Un conjunto de acciones conjuntas $\mathbf{a}^k \in \mathcal{A}$
- Una función de recompensa $R_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ que entrega a cada agente i una recompensa numérica r_i^k

$$\text{Donde } \mathbf{R}(\mathbf{s^k}, \mathbf{a^k}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{R_i}(\mathbf{s^k}, \mathbf{a^k})$$

Aprendizaje por refuerzo

Caso multiagente:

Teniendo en cuenta $Q(\mathbf{s}, \mathbf{a}) = \mathrm{E}\left\{R^k | \mathbf{s}^k = \mathbf{s}, \mathbf{a}^k = \mathbf{a}\right\}$

Es posible descomponer la función Q del sistema en una combinación lineal de funciones por agente:

$$Q(\mathbf{s}, \mathbf{a}) = \sum_{i=1}^{n} Q_i(\mathbf{s}_i, a_i)$$

Regla de actualización para el caso multiagente:

$$Q_i^{k+1}(\mathbf{s}_i^k, a_i^k) = (1 - \alpha^{k+1}) Q_i^k(\mathbf{s}_i^k, a_i^k) + \alpha^{k+1} \left[r_i^{k+1} + \gamma \max_{\mathbf{a}' \in \mathcal{A}} Q^k(\mathbf{s}^{k+1}, \mathbf{a}') \right]$$

Aprendizaje por refuerzo

Caso multiagente:

Teniendo en cuenta $Q(\mathbf{s}, \mathbf{a}) = \mathrm{E}\left\{R^k | \mathbf{s}^k = \mathbf{s}, \mathbf{a}^k = \mathbf{a}\right\}$

Es posible descomponer la función Q del sistema en una combinación lineal de funciones por agente:

$$Q(\mathbf{s}, \mathbf{a}) = \sum_{i=1}^{n} Q_i(\mathbf{s}_i, a_i)$$

Regla de actualización para el caso multiagente:

$$Q_i^{k+1}(\mathbf{s}_i^k, a_i^k) = (1 - \alpha^{k+1}) Q_i^k(\mathbf{s}_i^k, a_i^k) + \alpha^{k+1} \left[r_i^{k+1} + \gamma \max_{\mathbf{a}' \in \mathcal{A}} Q^k(\mathbf{s}^{k+1}, \mathbf{a}') \right]$$

Coordinación en MARL

Emerge la necesidad de coordinación de acciones entre agentes para maximizar la recompensa global a largo plazo

Problema de coordinación

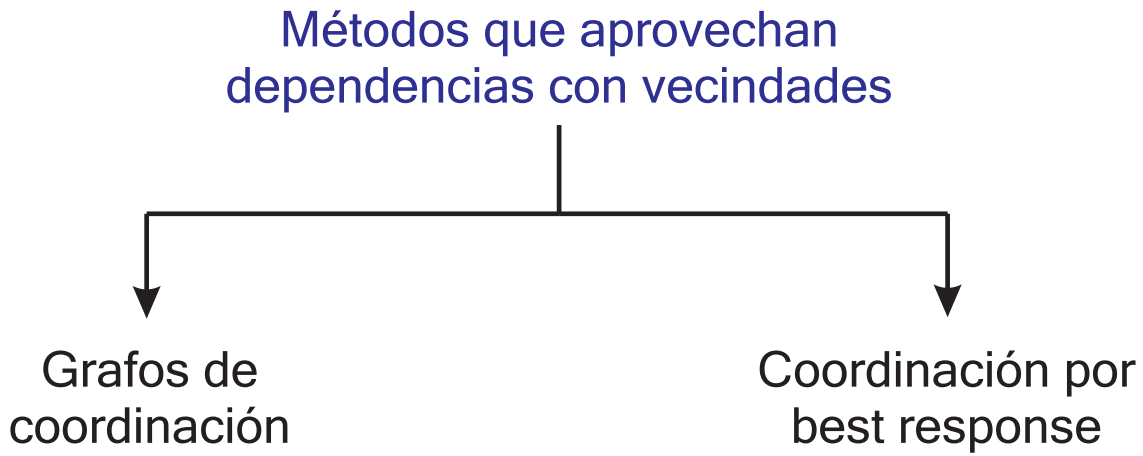
$$\text{Encontrar } \mathbf{a}' = \text{argmax}_{\mathbf{a}' \in \mathcal{A}} Q(\mathbf{s}, \mathbf{a}')$$

Enfoques para establecer coordinación

Principio de localidad

Dos objetos suficientemente alejados uno de otro no pueden influirse mutuamente de manera instantánea.

En el sistema de tránsito: la acción de cada agente influye en mayor medida en el estado percibido aredeor de su vecindad.



Método 1: Q-Learning y grafos de coordinación

- Representa, por medio de un grafo $G = (V, E)$, problemas en los cuales el agente i solo exhibe necesidad de coordinación con una vecindad $\Gamma(i)$.
- Permite la descomposición por arco de la función Q global.

$$Q(\mathbf{s}, \mathbf{a}) = \sum_{(i,j) \in E} Q_{ij}(\mathbf{s}_{ij}, a_i, a_j)$$

- Regla de actualización con Q-Learning[1]:

$$Q_{ij}^{k+1}(\mathbf{s}_{ij}^k, a_i^k, a_j^k) = (1-\alpha) Q_{ij}^k(\mathbf{s}_{ij}^k, a_i^k, a_j^k) + \alpha \left[\frac{r_i^{k+1}}{|\Gamma(i)|} + \frac{r_j^{k+1}}{|\Gamma(j)|} + \gamma Q_{ij}^k(\mathbf{s}_{ij}^{k+1}, a_i^*, a_j^*) \right]$$

[1] Propuesto por: J. Kok en *Cooperation and Learning in Cooperative Multiagent Systems*. Ph.D thesis, University of Amsterdam, 2006.

Método 1: Q-Learning y grafos de coordinación

- Representa, por medio de un grafo $G = (V, E)$, problemas en los cuales el agente i solo exhibe necesidad de coordinación con una vecindad $\Gamma(i)$.
- Permite la descomposición por arco de la función Q global.

$$Q(\mathbf{s}, \mathbf{a}) = \sum_{(i,j) \in E} Q_{ij}(\mathbf{s}_{ij}, a_i, a_j)$$

- Regla de actualización con Q-Learning[1]:

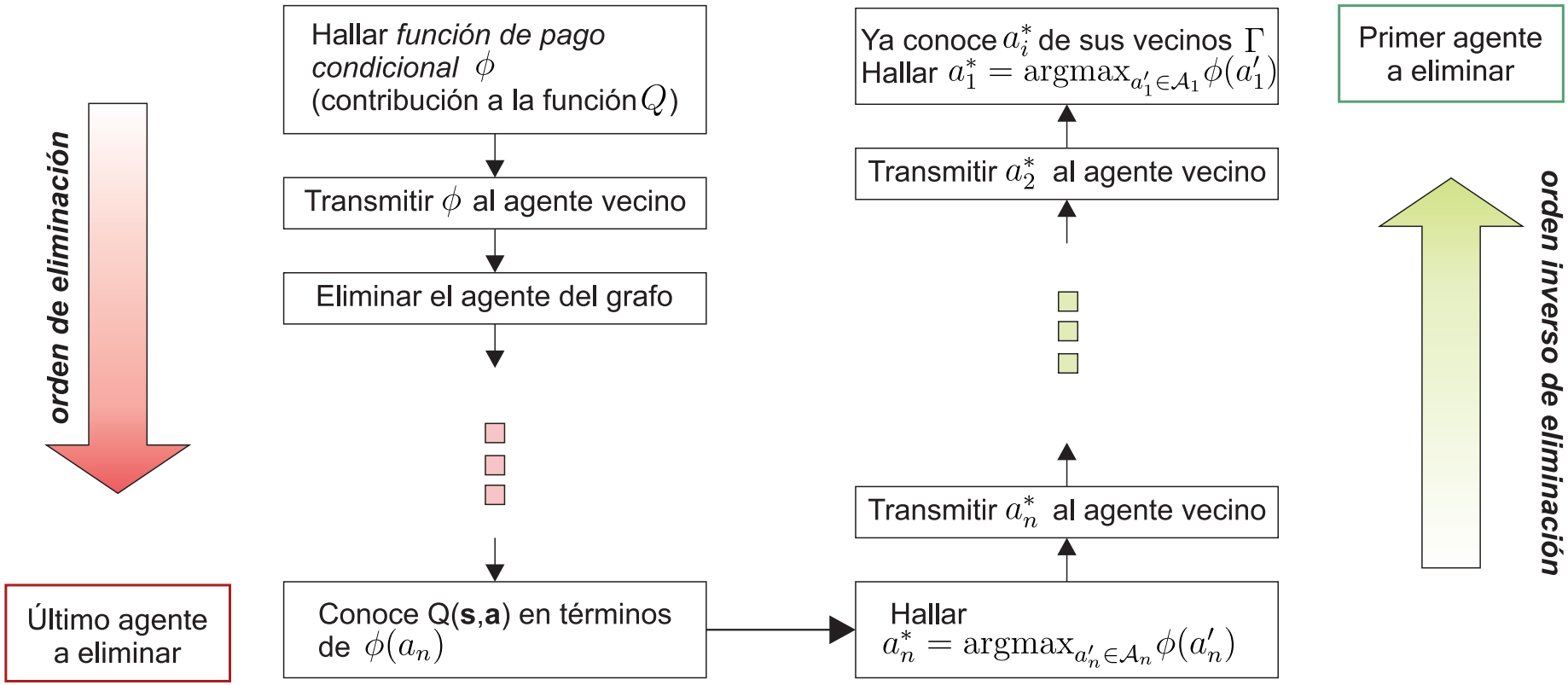
$$Q_{ij}^{k+1}(\mathbf{s}_{ij}^k, a_i^k, a_j^k) = (1-\alpha) Q_{ij}^k(\mathbf{s}_{ij}^k, a_i^k, a_j^k) + \alpha \left[\frac{r_i^{k+1}}{|\Gamma(i)|} + \frac{r_j^{k+1}}{|\Gamma(j)|} + \gamma Q_{ij}^k(\mathbf{s}_{ij}^{k+1}, a_i^*, a_j^*) \right]$$

$$a_i^*, a_j^* \in \underset{\mathbf{a}' \in \mathcal{A}}{\operatorname{argmax}}\, Q(\mathbf{s}, \mathbf{a}')$$

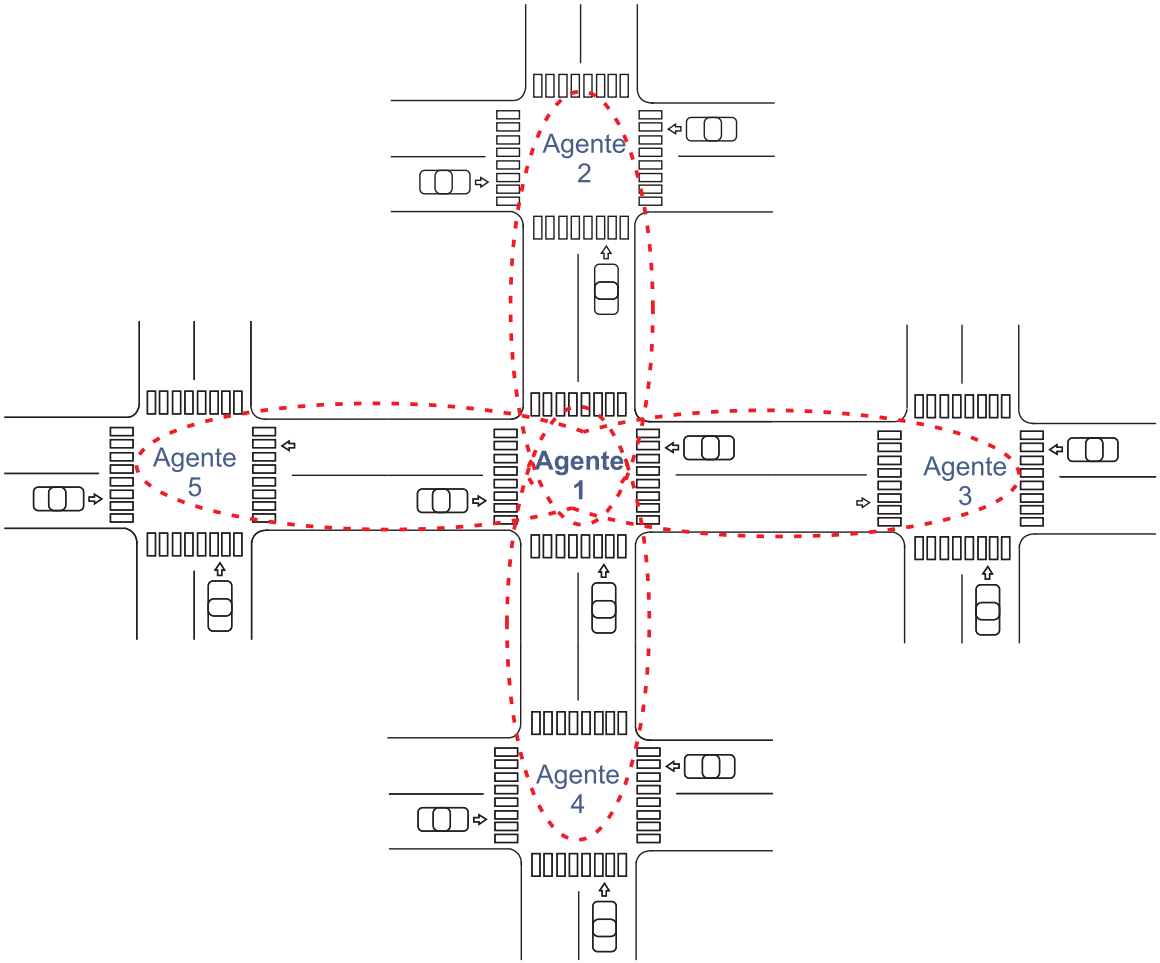
[1] Propuesto por: J. Kok en *Cooperation and Learning in Cooperative Multiagent Systems*. Ph.D thesis, University of Amsterdam, 2006.

Método 1: Q-Learning y grafos de coordinación

Algoritmo de eliminación de variable (VE): resuelve el problema de coordinación, encontrando $\mathbf{a}^* = \text{argmax}_{\mathbf{a}}\, Q(\mathbf{s}, \mathbf{a})$



Método 2: Q-Learning y *best response*



El agente i:

- Exhibe necesidad de coordinación con una vecindad NB_i
- Participa en un juego de dos jugadores con cada vecino $NB_i[j]$

Método 2: Q-Learning y *best response*

El agente i en cada periodo k:

1. Estima la política de sus vecinos:

$$\theta_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{\mathrm{NB}_i[j]}) = \frac{v(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{\mathrm{NB}_i[j]})}{\sum_{a_{\mathrm{NB}_i[j]} \in \mathcal{A}_{\mathrm{NB}_i[j]}} v(s^{k-1}_{i,\mathrm{NB}_i[j]}, a_{\mathrm{NB}_i[j]})}$$

Método 2: Q-Learning y *best response*

2. Actualiza los factores Q con cada vecino:

$$Q^{k}_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{i,\mathrm{NB}_i[j]}) = (1-\alpha)Q^{k-1}_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{i,\mathrm{NB}_i[j]}) + \alpha\left[r_i^k + \gamma \max_{a' \in \mathcal{A}} Q(s^k, a')\right]$$

Método 2: Q-Learning y *best response*

2. Actualiza los factores Q con cada vecino:

$$Q^k_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{i,\mathrm{NB}_i[j]}) = (1-\alpha)Q^{k-1}_{i,\mathrm{NB}_i[j]}(s^{k-1}_{i,\mathrm{NB}_i[j]}, a^{k-1}_{i,\mathrm{NB}_i[j]}) + \alpha\left[r^k_i + \gamma \mathrm{br}^k_i\right]$$

$$\mathrm{br}^k_i = \max_{a_i \in \mathcal{A}_i}\left[\sum_{a_{\mathrm{NB}_i[j]} \in \mathcal{A}_{\mathrm{NB}_i[j]}} Q_{i,\mathrm{NB}_i[j]}(s^k_{i,\mathrm{NB}_i[j]}, a_{i,\mathrm{NB}_i[j]}) \times \theta_{i,\mathrm{NB}_i[j]}(s^k_{i,\mathrm{NB}_i[j]}, a_{\mathrm{NB}_i[j]})\right]$$

Best response

- Función de pago $Q_i()$
- Estimativo θ_{-i} sobre las estrategias de los agentes vecinos

La estrategia $a_i \in \mathcal{A}_i$ para el jugador i es una *best response* si para todo a_i' se cumple:

$$Q_i(a_i, \theta_{-i}) \geq Q_i(a_i', \theta_{-i})$$

Método 2: Q-Learning y *best response*

3. Selecciona acción que corresponde a *best response* respecto a su vecindad:

$$a_i^k = \underset{a_i \in \mathcal{A}_i}{\text{argmax}} \left[\sum_{j \in \{1,2,\cdots,|\text{NB}_i|\}} \sum_{a_{\text{NB}_i[j]} \in \mathcal{A}_{\text{NB}_i[j]}} Q_{i,\text{NB}_i[j]}(s^k_{i,\text{NB}_i[j]}, [a_i \cup a_{\text{NB}_i[j]}]) \times \theta_{i,\text{NB}_i[j]}(s^k_{i,\text{NB}_i[j]}, a_{\text{NB}_i[j]}) \right]$$

[2] Basado en: El-Tantawy *et al.* en *Multiagent Reinforcement Learning for MARLIN-ATSC.* IEEE Transactions on Intelligent Transportation Systems, 2013.

Espacio de estados y acciones

Vector de estado

Estado de un agente de i accesos ($i \in \{\text{norte, este, sur, oeste}\}$), conformado por:

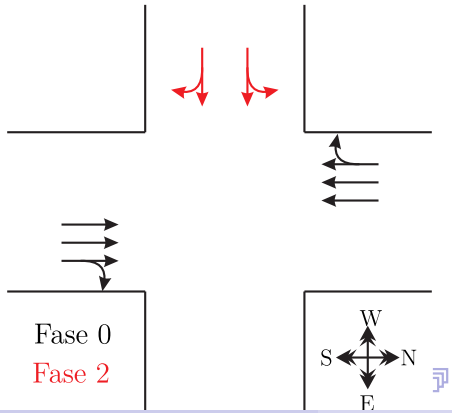
- Hora de día (h)
- Fase actual (k)
- Máxima longitud de cola (veh) por acceso (q_i)
- Tiempo de espera (min) de los vehículos por acceso (w_i)

} Discretizado por *Vector Quantization*

Acciones

Fase a aplicar con duración mínima:

- Q-VE: 20 segundos
- Q-BR: 14 segundos



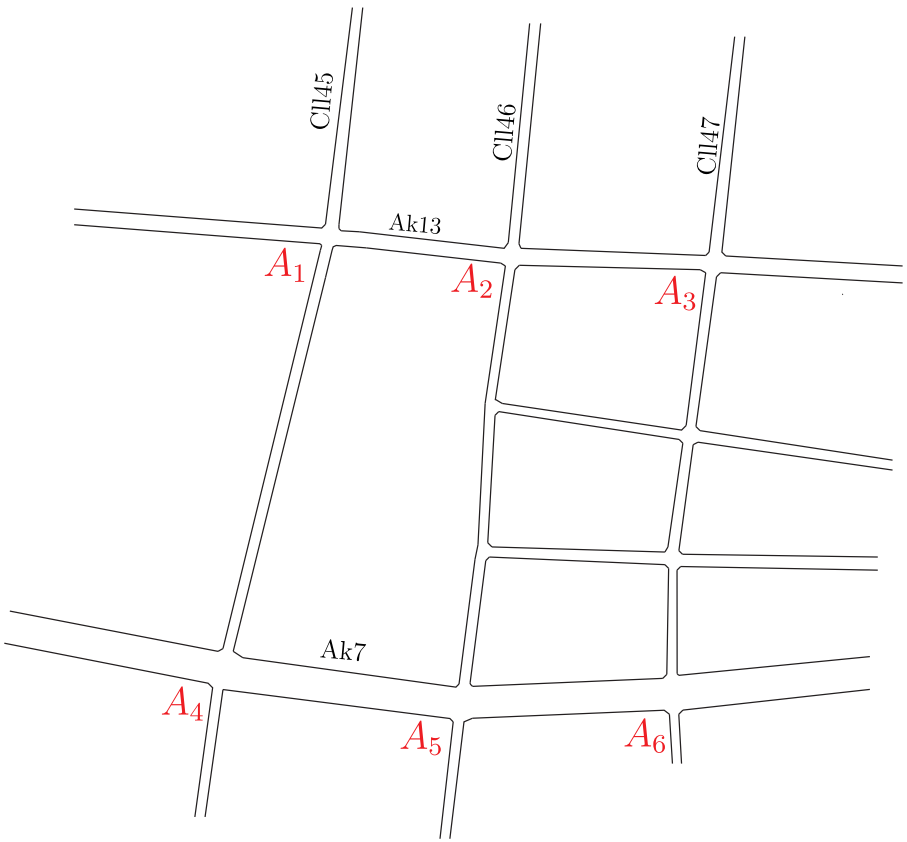
Función de recompensa para cada agente

$$r_i = -\sum_{k=1}^{N_i} \beta_q (q_k)^{\theta_q} + \beta_w (w_k)^{\theta_w}$$

Donde:

- N_i: número de accesos que tiene el agente i
- q_k: máxima longitud de cola del acceso k
- w_k: tiempo de espera de los vehículos en el acceso k
- β_q y β_w: coeficientes para el equilibrio de magnitudes de las variables q y w.
- θ_q y θ_w: términos potencia para equilibrar las longitudes de cola y tiempos de espera en los accessos.

Marco de prueba



Datos de flujos vehiculares y programas de fases de la Secretaría de Movilidad de Bogotá

Parámetros de simulación:

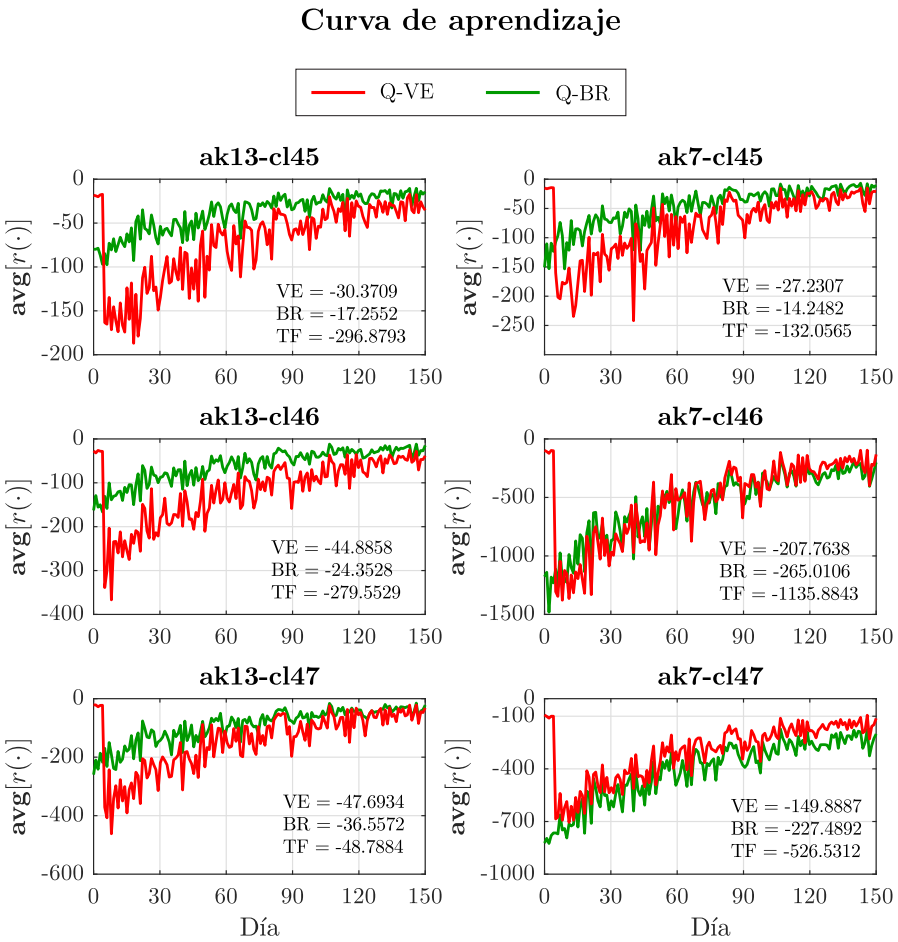
- Interacción con las agentes a través del simulador SUMO
- *Observación del sistema:* 5 episodios
- *Aprendizaje de política:* 150 episodios
- *Prueba de política:* 5 episodios

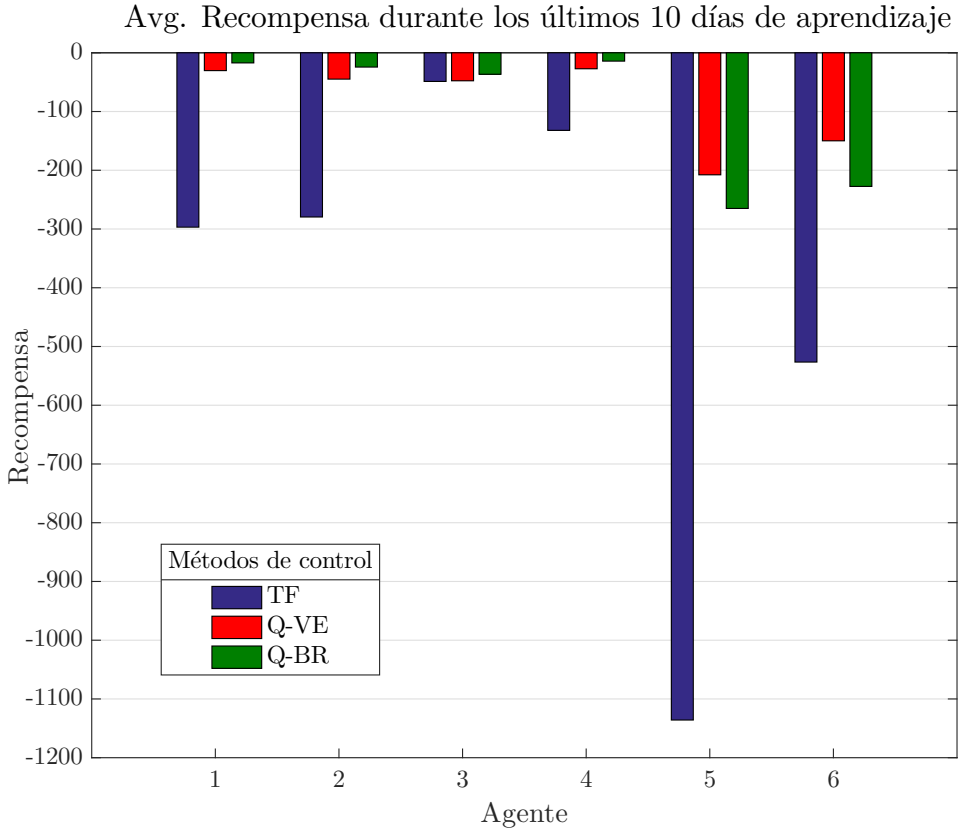
Detalles de ejecución:

- Entrenamiento en AWS
- Duración: 36 horas aprox.

Resultados

Curva de aprendizaje





Resultados

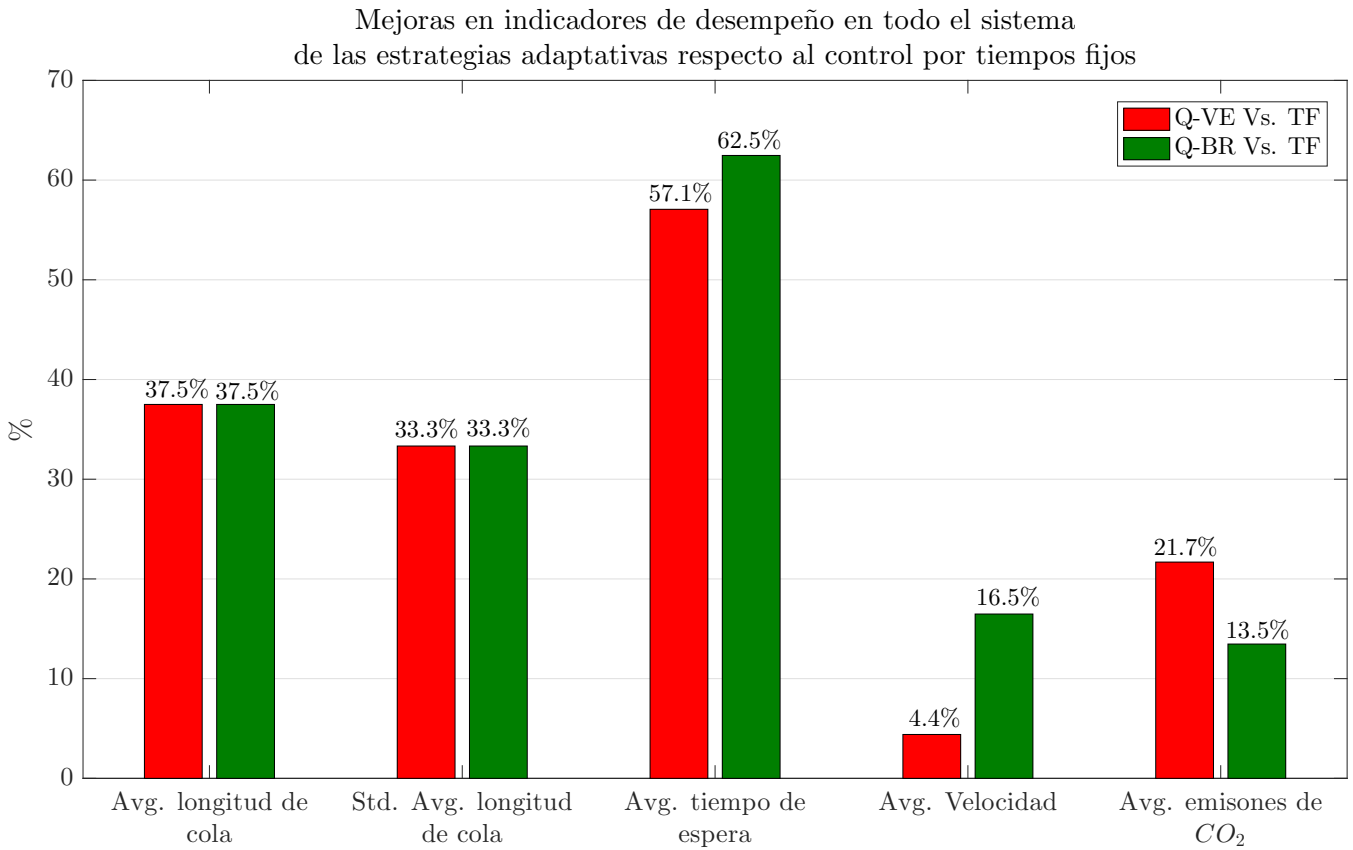
Indicadores de desempeño

- Máxima longitud de cola promedio por intersección (veh)
- Desviación estándar del promedio de las longitudes de cola en los accesos de las intersecciones (veh)
- Tiempo de espera promedio por vehículo (s/veh)
- Velocidad promedio (m/s)
- Emisiones promedio de CO_2 por intersección (mg)

Resultados

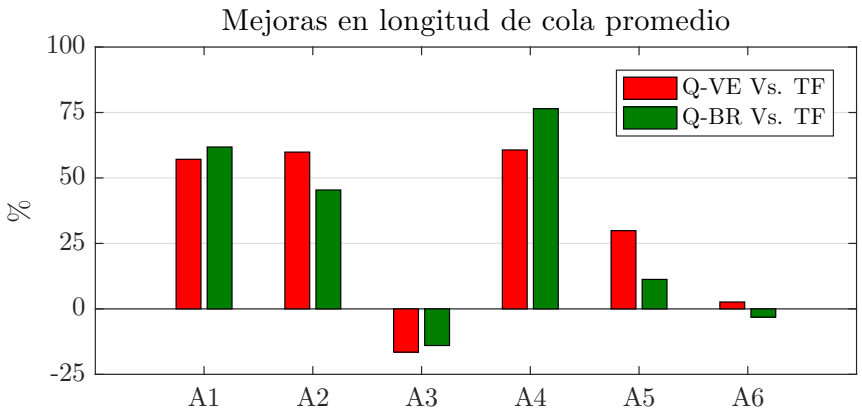
Indicadores de desempeño del sistema

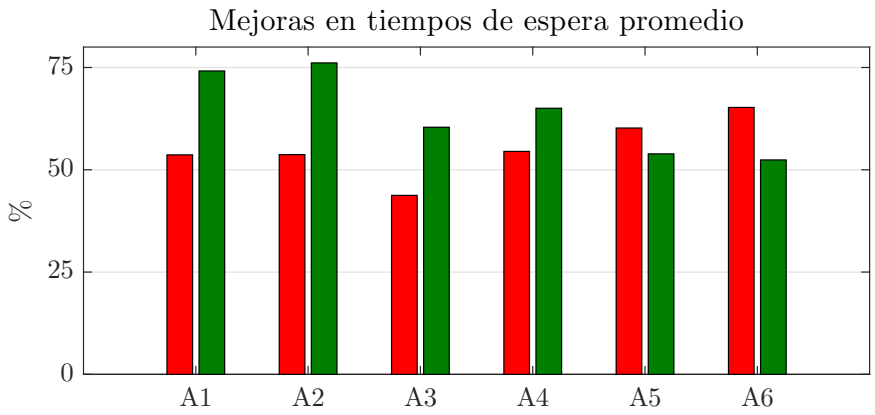
	Avg. longitud de cola (veh)	*Std. Avg. longitud de cola (veh)*	*Avg. tiempo de espera (s/veh)*	*Avg. velocidad promedio (Km/h)*	*Avg. emisiones de CO_2 (mg)*
TF	8	3	38,9	15,2	5140,9
Q-VE	5	2	16,7	15,9	4025,8
Q-BR	5	2	14,6	18,2	4448,3

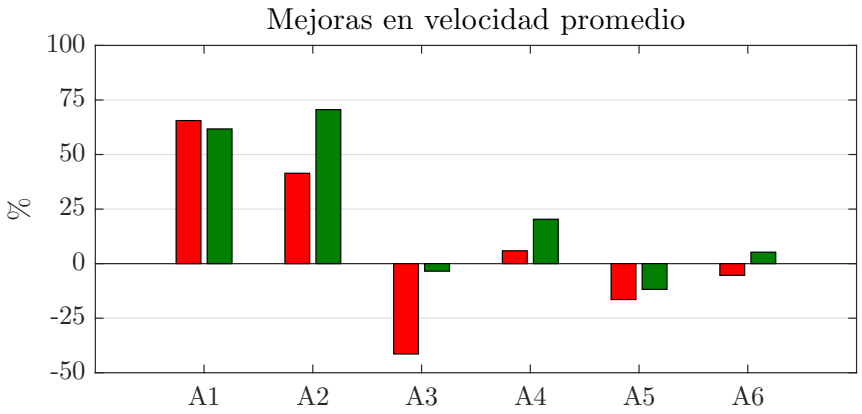


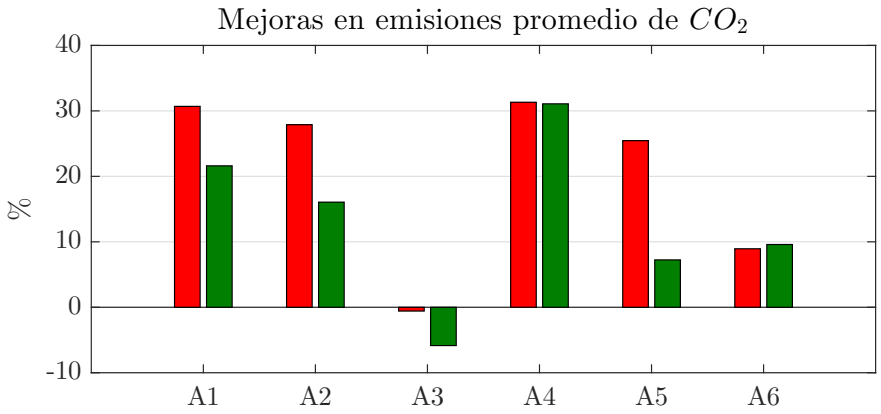
Resultados

Indicadores de desempeño por agente









Resultados

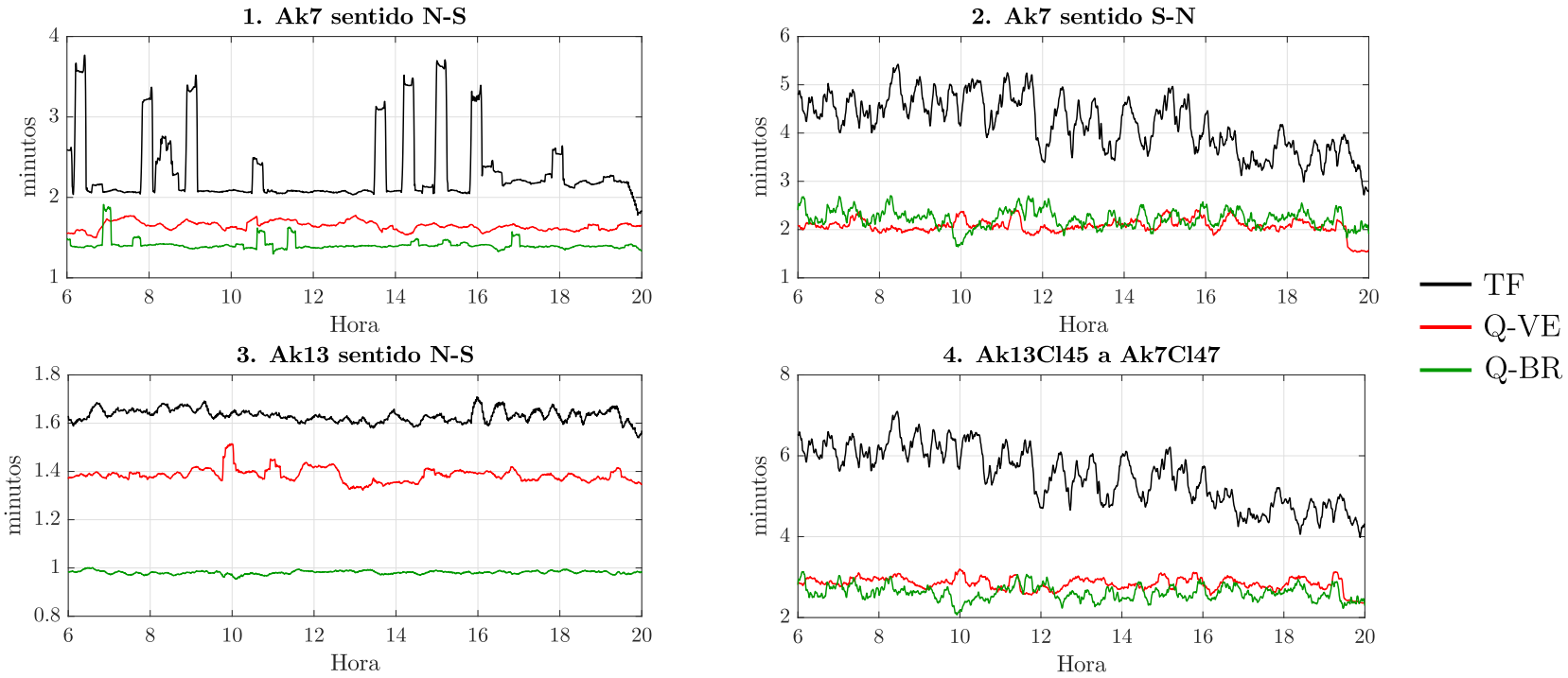
Tiempo de viaje para rutas principales

Tiempo de viaje promedio (min)

	Ruta 1	Ruta 2	Ruta 3	Ruta 4
TF	2.21	3.98	1.62	5.37
Q-VE	1.67	2.09	1.39	2.85
Q-BR	1.40	2.22	0.98	2.61

Mejoras en el tiempo de viaje promedio

	Ruta 1	Ruta 2	Ruta 3	Ruta 4
Q-VE Vs. TF	24,43 %	47,49 %	14,20 %	46,93 %
Q-BR Vs. TF	36,65 %	44,22 %	39,51 %	51,40 %



Conclusiones

- Las políticas aprendidas buscan priorizar el paso contínuo de grupos de movimientos, en lugar del paso de flujos paralelos.
- Para el sistema de tránsito, la aplicación del principio de localidad en la selección de las acciones de los agentes, conlleva a una política con mejor desempeño en los objetivos de minimización.
- El control adaptativo por medio de Q-BR presenta menor tiempo de viaje promedio y menores variaciones.

Conclusiones

- Las políticas aprendidas buscan priorizar el paso contínuo de grupos de movimientos, en lugar del paso de flujos paralelos.
- Para el sistema de tránsito, la aplicación del principio de localidad en la selección de las acciones de los agentes, conlleva a una política con mejor desempeño en los objetivos de minimización.
- El control adaptativo por medio de Q-BR presenta menor tiempo de viaje promedio y menores variaciones.

	Grafos de coordinación	*Best response*
Obtención de $\mathbf{a}^*$	Exacta por medio de VE	A nivel de vecindades
Escalabilidad	No facilmente	Completamente
Comunicaciones entre agentes	Sujetas a cambios	Definidas *a priori*
Observabilidad del sistema	Menor	Mayor

Trabajo futuro

1. Tener en cuenta otros objetivos en la función de recompensa (ej. velocidad promedio)
2. Tratar el problema con un espacio de estados continuo
3. Si se afronta el problema con un espacio de acciones continuo (ej. duraciones de las fases en una secuancia de fases fija), se podría incorporar el algoritmo de Consensus para la maximización de la función de recompensa

Videos del control adaptativo por Q-BR

Simulaciones

Hora pico 6am-7am

Hora valle 9am-10am

Hora pico 12m-1pm

¿Preguntas?