Repte DELOITTE: Report tècnic

Raúl Castillo Moraz, Pablo Smith Necpas Alba

November 9, 2025

Abstract

Aquest informe serveix com a suplement tècnic per a la presentació de PowerPoint, i presenta el desenvolupament i implementació d'un projecte d'anàlisi i disseny de noves línies de metro per a la ciutat de Barcelona, dins del marc del repte Deloitte. S'exposen les quatre línies principals del treball: recopilació i preprocesament de dades obertes, anàlisi exploratori amb càlcul de puntuacions esperades de recorregut, disseny de línies de metro i modelització de la demanda amb sèries temporals, integrant un xatbot per a consultes interactives. L'informe descriu detalladament les decisions tècniques adoptades, incloent la selecció de models de predicció (BATS), la representació relativa de les estacions per a l'optimització genètica i l'ús d'A* amb *sampling* per avaluar la qualitat de la xarxa. Els resultats obtinguts mostren millores significatives en el temps de recorregut mitjà i la reducció de desviaments, demostrant la viabilitat de l'enfocament integrat per planificar noves infraestructures de mobilitat urbana.

Contents

1 Introducció

1.1 Motivació

l projecte forma part del repte Deloitte "DataRide: Crea la nova línia de metro de Barcelona". L'objectiu és aplicar coneixements d'enginyeria i ciència de dades a un problema real de mobilitat urbana. El repte promou el treball en equip, el pensament analític, l'ús de dades obertes i la innovació tecnològica amb impacte social

2 Objectiu 1: Data collection and preprocessing

2.1 Font de Dades

S'han utilitzat conjunts de dades oberts provinents de:

- Open Data BCN: `https://opendata-ajuntament.barcelona.cat`
- ATM: `https://www.atm.cat/web/ca/dades-obertes`
- TMB: `https://www.tmb.cat/ca/transparencia/dades-obertes`
- INE i IDESCAT per a dades demogràfiques
- OpenStreetMap per a dades geoespacials

3 Objectiu 2: Exploratory data analysis (EDA)

3.1 Càlcul de la puntuació esperada amb A* i *sampling*

Per avaluar la qualitat d'una xarxa de metro, utilitzem una *funció de puntuació esperada* que estima el temps mitjà de recorregut d'una persona entre diferents barris de la ciutat. Aquesta puntuació combina la topologia de la xarxa de metro amb la població relativa dels barris, permetent prioritzar connexions que beneficiïn els usuaris reals.

3.1.1 Mètode

1. **Selecció de *landmarks*:** Es defineix un conjunt de punts d'interès (landmarks), corresponents als barris més rellevants de la ciutat. Per simular la probabilitat de desplaçament real, cada landmark rep un pes proporcional a la seva població.

2. **Sampling aleatori ponderat:** Degut a la gran quantitat de parells possibles de landmarks, seleccionem un subconjunt de parells de manera aleatòria, utilitzant els pesos per donar major probabilitat als barris més poblats. Això redueix la complexitat computacional sense perdre fidelitat en la representació de la demanda real.

3. **Càlcul del recorregut amb A*:** Per a cada parell de landmarks seleccionat, s'aplica l'algorisme A* sobre el graf de la xarxa de metro. L'heurística utilitzada és la distància euclidiana entre coordenades, que proporciona una aproximació efectiva del cost mínim esperat entre nodes. L'ús d'A* permet considerar tant els pesos del graf (temps de recorregut i espera) com les connexions reals existents.

4. **Combinació de resultats:** Les distàncies trobades es ponderen pel producte dels pesos dels dos landmarks implicats i es calcula la mitjana sobre tots els parells seleccionats. Així s'obté un *score* que reflecteix el temps mitjà esperat per un usuari qualsevol, ponderat segons la distribució poblacional.

5. **Normalització:** Per evitar que la puntuació depengui del nombre de parells processats, es divideix la suma ponderada entre el nombre de parells utilitzats en el càlcul.

3.1.2 Avantatges

- Permet comparar de manera objectiva diferents línies generades, tot incorporant tant la població dels barris com la distància real a recórrer.
- Redueix la complexitat computacional gràcies al *sampling* ponderat, sense sacrificar la representativitat dels desplaçaments.
- És flexible i pot incorporar fàcilment noves restriccions o pesos addicionals, com ara penalitzacions per temps d'espera o desviaments de ruta.

3.1.3 Codi Python simplificat

```
def expected_shortest_path_score(graph, positions, landmarks, weights):
    total_distance = 0
    pairs_processed = 0
    n = len(landmarks)

    for i in range(n):
        for j in range(i+1, n):
            if weights[i] > 0 and weights[j] > 0:
                d, _ = astar(graph, positions, landmarks[i], landmarks[j])
                if d < float('inf'):
                    total_distance += d * weights[i] * weights[j]
                    pairs_processed += 1

    if pairs_processed == 0:
        return float('inf')

    return total_distance / pairs_processed
```

Aquest procediment constituirá més endavant la base de la funció de fitness utilitzada en l'algorisme genètic per crear xarxes noves de metro.

3.2 Expansió d'una nova linia de metro a Horta.

A base de l'anàlisi exploratori de dades, vam veure que a Horta hi havia una possible deficiència i possibilitat d'expansió.

Vam afegir manualment les noves estacions a la base de dades i vam fer comparacions amb la distància mitjana per trajecte i també vam comparar la major diferència entre dos viatjes. Els resultats foren els següents:

Mètrica	Valor
Millora total	2,829%
Reducció màxima	324%

Table 1: Resum de resultats principals

Tot i que no hem tingut en compte els trajectes de bus per manca de temps, sí que hem considerat els desplaçaments a peu entre node i node. Aquest factor s'ha integrat en el càlcul del temps de recorregut esperat i es pot observar clarament com influeix en la configuració de la línia, especialment en la connexió entre barris propers on caminar és una alternativa viable i eficient. D'aquesta manera, es pot veure que realment seria un increment significatiu que s'hauria de sobmetre a estudi més rigorós.

4 Objectiu 3: System design and infrastructure

4.1 Algorisme genètic per a disseny de línies de metro

4.1.1 Plantejament del problema

L'objectiu és generar una nova línia de metro òptima que connecti diversos punts d'interès (landmarks) dins de la ciutat, minimitzant el temps de recorregut esperat entre els barris més poblats, i equilibrant factors com la rectitud de la línia, l'espai entre estacions i la longitud total de la trajectòria. Aquest és un problema de *optimització contínua* amb restriccions geogràfiques i costos associats a velocitats de recorregut i esperes mitjanes, on les funcions objectiu no són diferenciables ni lineals.

4.1.2 Raó per l'ús d'algorismes genètics

Hem optat per un *algorisme genètic* perquè:

- **Espai de solució contínua**: Les coordenades de les estacions poden variar en un pla geogràfic, per la qual cosa la solució és contínua i multidimensional.
- **Funció objectiu no diferenciable**: La heurística basada en el temps esperat de recorregut (A amb pesos) és discreta i depèn de la topologia de la xarxa, fent que mètodes basats en gradients no siguin aplicables.
- **Capacitat de explorar àmpliament l'espai**: Els algorismes genètics permeten explorar múltiples combinacions de línies simultàniament, mantenint diversitat i evitant minimitzacions locals.
- **Integració de múltiples criteris**: La funció de fitness combina conveniència (temps de recorregut), rectitud, homogeneïtat de separació i longitud total, permetent equilibrar objectius competitius.

4.1.3 Representació de les solucions i normalització

- **Representació relativa**: La primera estació es defineix amb coordenades absolutes, i les següents com desplaçaments respecte de l'anterior. Això facilita:
  - Mantenir la coherència geomètrica de la línia.

  - Aplicar operadors de creuament i mutació amb més estabilitat.

- **Normalització de pesos i coordenades**: Els pesos dels landmarks es normalitzen per assegurar que cada punt contribueixi proporcionalment al càlcul del temps esperat de recorregut, evitant que un barri molt poblós domini totalment la funció objectiu. Això garanteix una solució equilibrada.

4.1.4 Generació de la població inicial

La població inicial de línies es genera amb:

- Coordenades aleatòries dins d'un marge geogràfic de les estacions existents, assegurant cobertura realista.
- Desplaçaments aleatoris amb angles lleugerament variats per crear corbes suaus i evitar trajectòries artificials.

4.1.5 Funció de fitness

La funció de fitness combina diversos components ponderats:

- **Score de recorregut esperat**: La puntuació explicada a la secció anterior, que fa servir A* i sampling per estimar la durada del recorregut promig d'una persona.
- **Rectitud de la línia**: Penalitza corbes brusques i canvis d'angle elevats entre estacions consecutives.
- **Homogeneïtat d'espaiament**: Penalitza distàncies molt desiguals entre estacions, garantint una distribució uniforme.
- **Longitud total**: Evita línies excessivament llargues, mantenint costos d'inversió raonables.

4.1.6 Creuament i mutació

- **Creuament BLX-alpha sobre desplaçaments**: Permet combinar les característiques de dues línies pares generant nous individus que exploren l'espai de solució de manera contínua.
- **Mutació amb soroll gaussià decaient**: Introduïm petites variacions amb magnitud que decreix amb les generacions (*annealing*) per estabilitzar la recerca cap a solucions òptimes, equilibrant exploració i explotació.

4.1.7 Avaluació i selecció

- **Torneig per selecció de pares**: Selecciona individus amb alta fitness per creuar-los, mantenint la pressió selectiva però preservant diversitat.
- **Elitisme**: Manté els millors individus de cada generació sense modificar-los, assegurant que el progrés acumulat no es perdi.

4.1.8 Resultat final

El resultat és una línia de metro representada amb coordenades relatives, que després es poden convertir a absolutes per visualització i implementació. Aquesta línia és òptima segons els criteris combinats de conveniència, rectitud, homogeneïtat i longitud, i es genera de manera robusta gràcies a la naturalesa adaptativa dels algorismes genètics.

5 Objectiu 4: Algorisme de predicció i Xatbot

Aquesta secció descriu l'algoritme de predicció de la demanda per estació i el disseny d'un chatbot que respon a les consultes dels usuaris combinant el reconeixement d'entitats basat en NER amb el reconeixement d'intencions manual, i afegint-hi LLMs (models grans de llenguage) especialitzats en català.

5.1 Algorisme de predicció

5.1.1 Plantejament del problema

L'objectiu és predir la demanda a cadascuna de les estacions basant-nos en dades històriques.

5.1.2 Models candidats

- **Models de sèries temporals**: ARIMA, Prophet, LSTM/GRU, BATS. Cadascun d'aquests models aporta avantatges específics: ARIMA és molt efectiu per capturar patrons lineals i autocorrelacions; Prophet facilita la incorporació d'estacionalitats múltiples i dies festius; LSTM/GRU poden modelar dependències a llarg termini en dades seqüencials no lineals; i BATS combina transformacions Box-Cox, modelització de residus ARIMA, tendència i estacionalitat.
- **Enfocaments híbrids**: models d'aprenentatge automàtic basats en característiques (features) derivades de retardos i finestres temporals. Aquests mètodes permeten combinar informació seqüencial amb atributs addicionals (per exemple, condicions meteorològiques, esdeveniments especials) i sovint milloren el rendiment quan hi ha factors externs influents.

5.1.3 Pipeline d'entrenament (nivell alt)

1. Preparar el conjunt de dades amb divisió temporal en train/validation/test. En concret, vam fer servir l'any sencer de 2024 com a test. És fonamental respectar l'ordre temporal per evitar fuites d'informació cap al futur i assegurar una avaluació realista del model.
2. Escalat i codificació de característiques. La normalització o estandardització de les variables contínues i la codificació de les categòriques són passos crítics per a models basats en ML i també poden estabilitzar alguns models de sèries temporals.
3. Entrenament dels models candidats i validació utilitzant les mètriques seleccionades (RMSE, MAE, AUC, etc.). La selecció de la mètrica depèn de la naturalesa de la variable objectiu: RMSE i MAE per a prediccions contínues, AUC per a classificació binària o probabilitats d'incident.

4. Selecció del millor model i producció d'una avaluació final, incloent gràfics de calibració i comparació de prediccions vs valors reals. Això permet visualitzar tant l'ajust del model com la seva capacitat per generalitzar a noves dades.

Aquest pipeline assegura una aproximació sistemàtica i robusta per identificar el model més adequat, minimitzant sobreajustaments i maximitzant la precisió de les prediccions per a la planificació i gestió dels recursos futurs.

5.1.4 Model final

Finalment, vam decidir optar pel model **BATS** (Box-Cox, ARIMA residuals, Trend and Seasonality), ja que va oferir resultats molt satisfactoris i representava un bon compromís entre eficiència computacional i rendiment predictiu. La decisió no va ser arbitrària: vam avaluar diversos models de sèries temporals, incloent-hi ARIMA simple, Holt-Winters i models basats en xarxes neuronals. Tot i que alguns models més complexos podrien haver proporcionat lleugerament millors mètriques d'error, el BATS ens va permetre capturar tant la complexitat estacional com les possibles transformacions de variància (mitjançant Box-Cox) sense sobreajustar el model.

Un dels punts clau que ens va portar a escollir BATS és la seva capacitat de modelar residus ARIMA després de tractar la tendència i l'estacionalitat. Això ens permet tractar amb precisió els components més subtils de la sèrie temporal que un model lineal simple no podria capturar. A més, la inclusió de la transformació Box-Cox ajuda a estabilitzar la variància de la sèrie, especialment quan les dades tenen valors extrems o oscil · lacions fortes, millorant així la precisió de les prediccions futures.

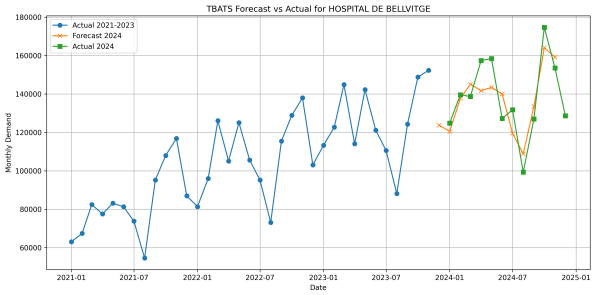

Figure 1: Resultats per l'estació de HOSPITAL DE BELLVITGE

Com es pot veure en la figura anterior, el model ofereix un ajust molt satisfactori a les dades històriques. Aquest rendiment suggereix que el BATS podria ser una eina molt útil per predir la demanda futura de les estacions, permetent planificar millor els recursos i escalar serveis segons les necessitats futures.

5.2 Disseny del xatbot

El xatbot funciona com una interfície de conversa per consultar informació sobre les estacions i els resultats de les prediccions. Segueix el següent flux de procés:

1. Preprocessament: tokenització, conversió a minúscules i normalització simple

2. Reconeixement d'entitats (NER): identificació dels noms d'estació, identificadors i altres entitats

3. Reconeixement d'intencions: sistema basat en regles fetes a mà que associa la formulació de l'usuari amb les intencions

4. Gestor de diàleg: dirigeix la consulta a l'obtenció de dades, a les prediccions o a la sol · licitud de clarificació

5. Generació de resposta: respostes basades en plantilles, completades amb les dades corresponents

5.2.1 NER (reconeixement d'entitats)

Vam utilitzar un pas dedicat de NER per detectar les entitats corresponents a les estacions. Les opcions inclouen:

- Models lleugers (spaCy, flair) entrenats o afinats amb els noms de les estacions
- Reconeixedor basat en *gazetteer* o llistat (ràpid i determinista) poblats amb la llista d'estacions

5.2.2 Reconeixement d'intencions

El reconeixement d'intencions és molt més senzill: el sistema només processa tres tipus de consultes principals en llenguatge natural:

- **Com arribar d'una estació a una altra**: preguntes del tipus "Com puc anar de Sants a Universitat?".
- **Informació d'una estació concreta**: preguntes com "Quina informació tens sobre l'estació de Diagonal?".
- **Informació d'una línia de metro**: per exemple, "Dóna'm dades de la Línia 3".

El reconeixement es fa mitjançant un conjunt de regles senzilles que detecten patrons bàsics i entitats identificades pel mòdul NER. subsubsectionGestió de consultes complexes amb AIna Tot i que el sistema de reconeixement d'intencions és lleuger i basat en regles simples, hi ha casos en què la consulta de l'usuari és massa ambigua o complexa per ser processada correctament. Per aquestes situacions, s'integra un mecanisme de suport que utilitza **AIna**, la intel · ligència artificial oberta catalana, com a sistema de *fallback*.

Quan el reconeixedor d'intencions retorna una intenció desconeguda o incerta, el text de la consulta s'envia a AIna per obtenir una resposta contextualitzada en català, mantenint la coherència amb el domini del metro de Barcelona.

Listing 1: Integració amb AIna com a fallback

```
if intent == 'DESCONEGUDA':
    resposta = AIna.generate_response(text)
else:
    resposta = handle_intent(intent, entities)
```

Aquesta integració permet al xatbot mantenir la fluïdesa de la conversa i oferir una experiència més natural, tot aprofitant els recursos lingüístics i culturals propis del català.

6 Conclusions i passos a seguir

6.1 Conclusions principals

- Hem desenvolupat un procés complet per explorar i dissenyar noves línies de metro a Barcelona, integrant anàlisi de dades obertes, modelització de sèries temporals i algoritmes genètics per optimitzar la ubicació de noves estacions.
- La funció de puntuació esperada basada en A* i *sampling* permet avaluar de manera realista el temps mitjà de recorregut dels usuaris, tenint en compte tant els desplaçaments a peu com la població dels barris.
- L'algorisme genètic ha demostrat ser una eina efectiva per generar línies òptimes, equilibrant diversos criteris: conveniència, rectitud, homogeneïtat d'espaiament i longitud total.
- El model de predicció BATS proporciona prediccions fiables de la demanda futura per estació, facilitant la planificació dels recursos i la gestió de la capacitat.
- El xatbot desenvolupat combina reconeixement d'entitats (NER) i regles de reconeixement d'intencions amb un sistema de fallback basat en AIna, permetent una interfície interactiva en català que respon de manera coherent i contextualitzada.

6.2 Passos a seguir

- **Validació i refinament de la xarxa de metro generada:** realitzar estudis més detallats amb dades de mobilitat reals i considerar altres modes de transport com autobusos i trens.
- **Escalabilitat de computació:** Amb ordinadors més potents o entorns de computació distribuïda, es podrien executar més generacions de l'algorisme genètic, augmentar el nombre de *sampling* per a la puntuació esperada i explorar un espai de solució més ampli, millorant així la qualitat i robustesa de les línies generades.
- **Optimització del xatbot:** entrenar models NER més precisos i ampliar el reconeixement d'intencions per cobrir més tipus de consultes.
- **Automatització i escalabilitat:** implementar pipelines d'ML Ops per facilitar l'actualització periòdica de les prediccions i la reavaluació de la xarxa.
- **Estudis d'impacte socioeconòmic i sostenibilitat:** avaluar els efectes de les noves línies sobre la mobilitat, accessibilitat i emissions, per prioritzar solucions sostenibles.

Moltes gràcies per llegir el nostre report.