Description for the Public Transport GTFS Dataset

The dataset contains the historical information of buses on the high-frequency inner-city bus line 4 in Stockholm, Sweden, shown in Fig. 1. Bus line 4 traverses the city centre and is the busiest bus line in Stockholm, with buses departing every 4-6 minutes between 06:50 to 19:00 on weekdays. Spanning a distance of approximately 12.4 kilometres, bus line 4 comprises a total of 28 stops in the northbound (Gullmarsplan-Radiohuset) directions. It takes around 60 minutes for a typical vehicle to travel the entire route in one direction. For the purpose of analysis, the dataset along the selected route and direction between 6:00 a.m. and 10:00 p.m. from January to June 2022 was collected.

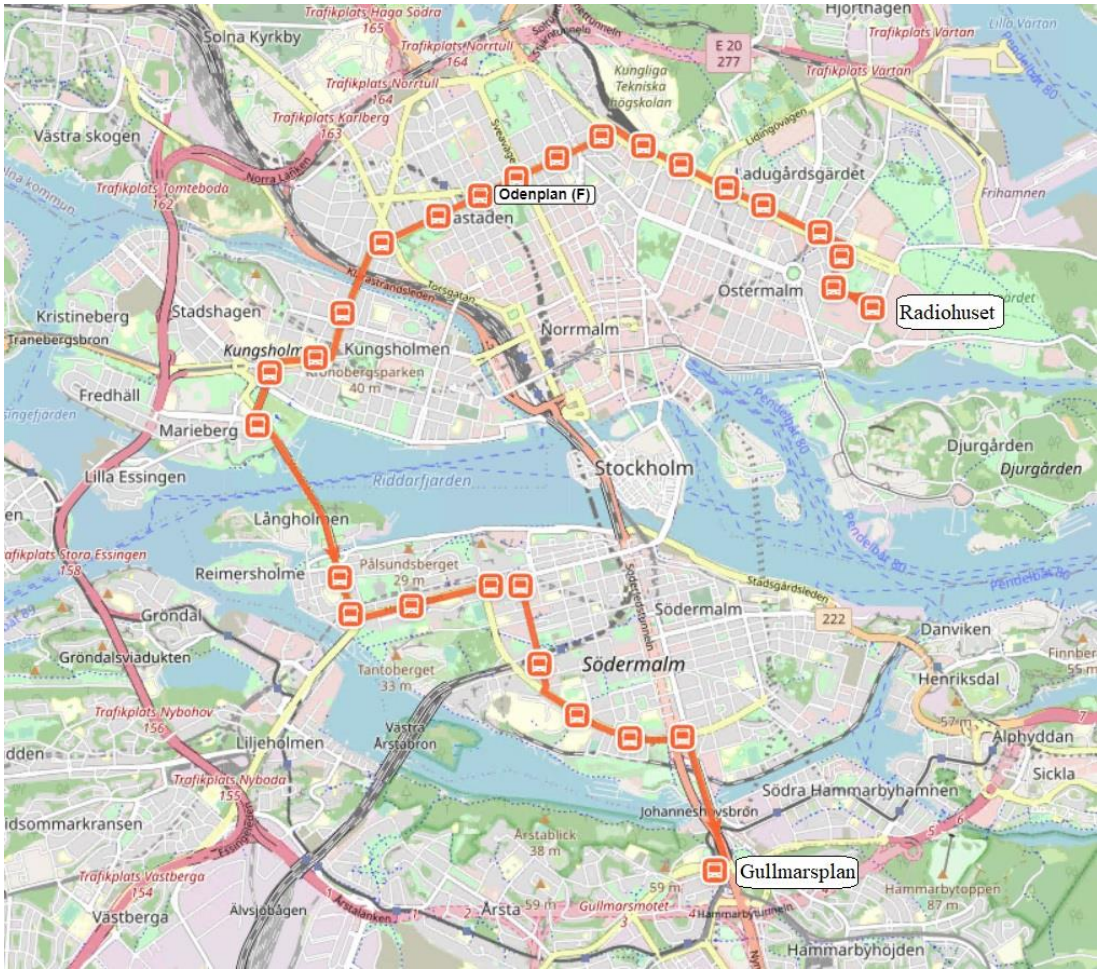

Fig. 1 The route of bus line 4, Stockholm. Source: OpenStreetMap.

The dataset includes features:

- Basic information (Calendar date, route ID, stop ID, etc.).
- Bus operation information (dwell time, upstream stop delays, scheduled travel time, origin delays, etc.)
- Categorical variables (day of week, time of day, weather, temperature).

Field name	Sub_fields	Description
Basic information	*Calendar_date*	The date of this trip instance in YYYMMDD format.
	route_id	The route_id from the GTFS that this selector refers to.
	bus_id	Indicates the id of current bus/vehicle.
	stop_sequence	Indicates the sequence of the current stop.
Bus operation information (continuous variables)	*arrival_delay*	The arrival delay of bus j at stop i, that is the difference between the actual arrival time and the scheduled arrival time of bus j at stop i.
	dwell_time	Actual dwell time at the consecutive upstream stop, that is the difference between the actual departure time and the actual arrival time of bus j at the consecutive upstream stop i – 1. **Note:** There is a significant and positive correlation between current stop's arrival delays and the actual dwell time at the consecutive upstream stop i – 1.
	travel_time_for_previous_section	Actual running time between stop i–2 and i-1, that is the difference between the actual arrival time at stop i-1 and stop i – 2. **Note:** The travel time at the previous section has a slightly positive effect on the arrival dela at the next stop.
	scheduled_travel_time	Scheduled running time between stop i–1 and i, that is the difference between scheduled arrival time at stop i and the scheduled departure time at station i – 1. **Note:** There is a significant negative correlation between the current stop's arrival delays and the scheduled travel time for the current section (e.g., the section between stop i-1 and stop i).
	upstream_stop_delay	Actual arrival delay of bus j at upstream stops i-1, that is the difference between the actual and scheduled arrvial time at upstream stop i-1.

		Note: It has a notable positive effect on current stop's arrival delays, as they can directly propagate to the current station, exacerbating the delay.
	origin_delay	Actual arrival delays at origin stop, that is the actural arrival delays of bus k at origin stop. **Note:** Delays at origin stop have a slight positive impact on the arrival delays of downstream stops.
	previous_bus_delay	Actual arrival delay of the bus j-1 before bus j at stop i, that is the difference between the actual and scheduled arrival time of previous bus j-1 at stop i. **Note:** Delays of preceding buses have a statistically positive impact on current bus delays.
	previous_trip_travel_time	Actual running time of the preceding bus j-1 between stop i-1 and i. **Note:** The travel time for the previous trip has a tiny positive effect on the current trip arrival delay.
	traffic condition	Current traffic condition, that is the historical mean for arrival delays of bus j at stop i during the same hour interval. **Note:** The current traffic conditions can contribute to current arrival delays.
	recurrent_delay	Recurrent delays, that is the historical mean for the travel time of bus j at stop i during the same hour in the same weekdays. It captures the recurring bus delays observed at current stop during this time periods. **Note:** The recurrent delay is reflective of inherent factors that impact traffic flow and conditions during specific times of the day. Therefore, higher levels of recurrent delays result in longer travel times and increased arrival delays at next stop.
Categorical variables	*day_of_week*	Weekdays: [Monday, Friday]; Weekends: [Saturday, Sunday]
	time_of_day	Morning peak: [6am, 9 am]; Afternoon peak: [4pm, 7pm]; else Off-peak

	temperature	Normal temperature: Temperature > 0◦C; Cold: Temperature [−5◦C, 0◦C]; Extra cold: Temperature < −5◦C;
	weather	Normal weather: No rain or snow; Light rain: Precipitate [0mm, 10mm]; Rain: Precipitate > 10mm; light snow: snow depth [0mm,10mm];Snow:snow depth > 10mm;

Note: The continuous variables units are in seconds.

To facilitate data analysis, the categorical variables were transformed into dummy variables (i.e., factor(weather) Light_Rain, factor(weather) Light_Snow, factor(weather) Normal, factor(weather) Rain, etc.) using the one-hot encoding.

Please contact Qi Zhang at qzhan@kth.se or Zhenliang Ma (zhema@kth.se) if you may find any problems or have any questions about the dataset.