

渋滞予測データチャレンジ仕様書 (Ver.1.0)

NEXCO東日本・東京大学大学院情報学環

データチャレンジ仕様書の構成

1. 本ドキュメントの目的及びデータの利用について
2. データの概要について
3. 予測対象の高速道路区間について（関越道・館山道）
4. 高速料金・ルート検索のデータ（ドラぷらルート検索データ）について
5. 使用データの仕様について
6. 予測対象と予測タスクについて
7. 提出データの仕様について
8. 評価基準について

1. 本ドキュメントの目的及び提供データの利用について

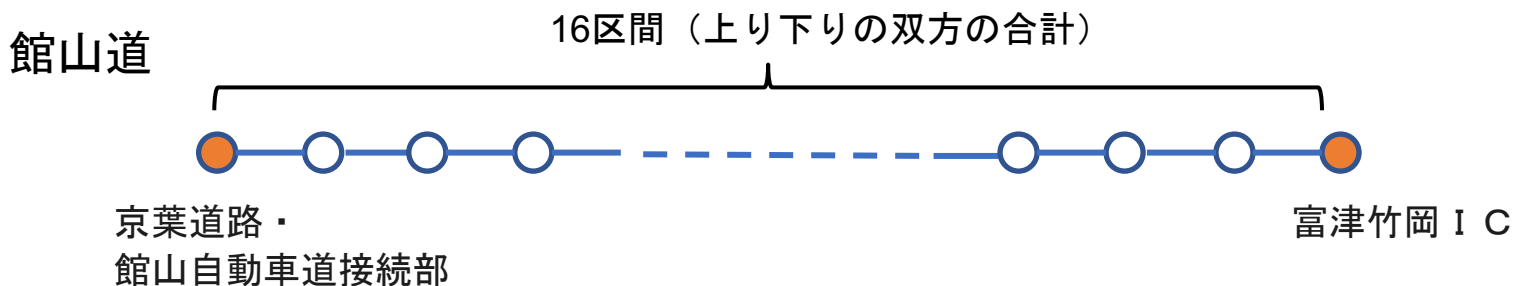
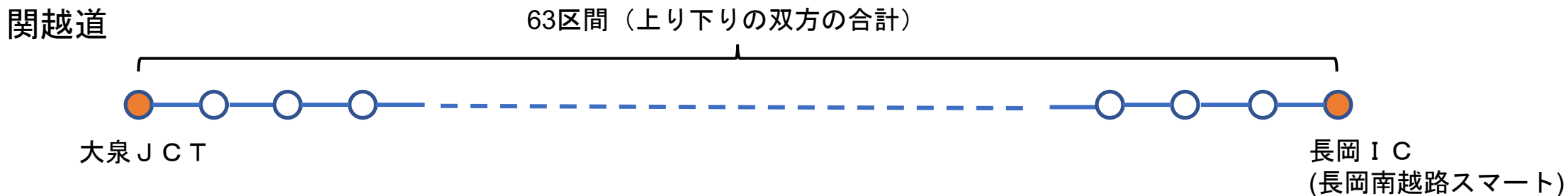
- 本ドキュメントは、NEXCO東日本・東京大学大学院情報学環が共催する渋滞予測データチャレンジ（以降「データチャレンジ」という。）で使用するデータの仕様について記したものです。
- データチャレンジにおいて提供されるデータは、本データチャレンジに参加する目的においてのみ使用可能であり、本目的以外での使用を禁止します。詳細は、データチャレンジチャレンジ規約を参照してください。

2. データの概要について

- データチャレンジにおいては、過去の交通量に関するデータ（以下、トラカンデータ）と高速料金・ルート検索のデータ（以下、ドラぷらルート検索データ）から、未来の高速道路（関越道・館山道）の各区間（上り下りの双方）における渋滞の有無を予測します。
- データチャレンジにおいて提供されるデータは、次の3種類です。
 - ▶ 高速道路の構造（区間）について示したデータ（道路構造データ）
 - ◆ 高速道路の区間を挟むIC（インターチェンジ：高速道路の出入口）の名称、位置や当該区間に設置されたトラフィックカウンターの位置等のデータ
 - ▶ トラカンデータ
 - ◆ 高速道路に設置された計測機。詳細は、下記を参照。
<https://www.driveplaza.com/safetydrive/mamechishiki/029.html>
 - ▶ ドラぷらルート検索データ
 - ◆ 4章及び下記ドラぷらのウェブサイトを参照。
<https://www.driveplaza.com/dp/SearchTop>

3.予測対象の高速道路区間について

- データチャレンジにおいては、関越道の「大泉JCT」から「長岡IC（長岡南越路スマート）」（区間数63区間（上り下りの双方の合計））、館山道の「京葉道路・館山自動車道接続部」から「富津竹岡IC」（区間数16区間（上り下りの双方の合計））を渋滞予測の対象とします。



4. ドラぷらルート検索データについて

- データチャレンジにおいては、コンテスタントの利便性のため、ドラぷらの検索履歴の生データではなく、検索履歴データを、時刻、高速道路の区間及びトラカンと関連した形で提供を行います。その手法について、この章で説明します。

4.1 「ドラぷら 高速料金・ルート検索サービス」について

■ 検索の流れ

- ▶ ルートの出発・到着インターチェンジ（IC）を指定する
 - ◆ ルート中の経由ICを指定可能（本データチャレンジでは考慮しない）
 - ◆ 検索するルートを使用する予定日時を指定可能
 - ◆ 指定した日時が出発時なのか到着時なのかを指定可能

■ ドラぷら 料金・ルート検索1件ごとにログ（以降「ドラぷら 検索ログ」）が1行記録される（以下5つの項目）

- ▶ 検索日時
 - ◆ 料金・ルート検索が実行された日時
- ▶ 出発IC
 - ◆ 高速道路に入るIC
- ▶ 到着IC
 - ◆ 高速道路から出るIC
- ▶ 指定日時
 - ◆ 検索されたルートを移動する予定の日時
 - ◆ 指定されない場合はnullがセットされる
- ▶ 指定種別
 - ◆ 指定日時が出発時か到着時か
 - ◆ 指定されない場合はnullがセットされる

ドラぷら 高速料金・ルート検索サービス トップページ

<https://www.driveplaza.com/dp/SearchTop>

「地図からIC名を調べて入力」または「住所やキーワードなどから入力」をクリックすると「地図」「住所・キーワード」「道路名」「履歴」いずれかの方法でIC名を調べられます。

出発IC
練馬 × 地図からIC名を調べて入力
+ 経由地を追加する
到着IC
湯沢(関越道) × 地図からIC名を調べて入力
住所やキーワードなどから入力

日時 2023/03/01 10 時 00 分
● 出発時 ○ 到着時

詳細条件を設定する

検索する

ルートの出発・到着ICを指定

ルートを使う予定の日時を指定
(指定しない場合は null が入る)

検索日時	出発IC	到着IC	指定日時	指定種別
2023/mm/dd HH:MM	練馬	湯沢	2023/3/1 10:00	出発

4.2 ドラぷら検索ログの構造化

- 本データチャレンジでは、コンテスタントの利便性のために、トラカンデータと関連する形でドラぷら検索ログの構造化を行い、構造化後のドラぷら検索ログを「ドラぷらルート検索データ」として提供する
- ドラぷら検索ログは渋滞予測の入力データとして活用しづらいため構造化する
 - ▶ ドラぷら検索ログからは「検索されたルートがどの区間を通過するのか」という詳細な経路情報や「その区間をいつ通過するのか」という時間情報を直接取り出すことができない
- 構造化の際、ドラぷら検索ログを2種類に分割
 1. 時間指定あり検索ログ: ドラぷらの料金・ルート検索時に移動日時を指定しているログ
 - ◆ より直接的に将来の潜在的な交通量として解釈できる
 2. 時間指定なし検索ログ: ドラぷらの料金・ルート検索時に移動日時を指定していないログ (= 指定日時が null)
 - ◆ 将来のどのタイミングで交通状況に影響を及ぼすかは分からないが、将来の交通状況と何らかの相関が存在すると仮定
- 時間指定あり検索ログ、時間指定なし検索ログを別の方法で構造化し、その結果の値を「時間指定あり検索数」「時間指定なし検索数」としてそれぞれ定義する (4.3及び4.4を参照)

4.3 ドラブラ検索ログ構造化手順: 時間指定あり検索ログ

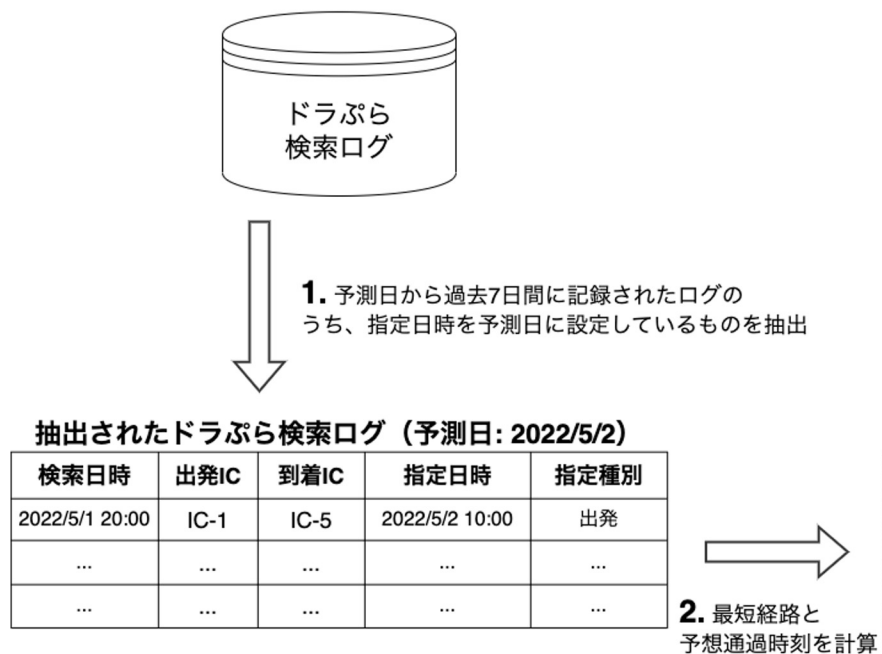
- 時間指定あり検索ログを以下の手順で時間指定あり検索数テーブルの形に構造化する

1. 予測対象日から過去7日間に記録されたドラぷら検索ログのうち、指定日時が予測対象日に設定されているログのみを抽出

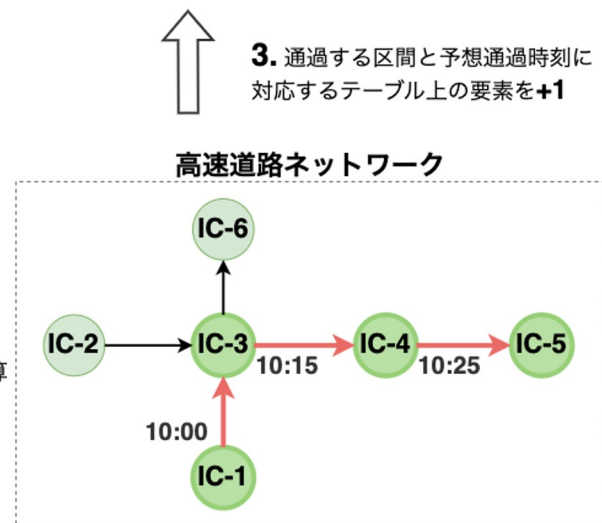
2. 抽出されたログの出発ICから到着ICに至る最短経路と、経路中の各区間の予想通過時刻を計算

- ▶ 各区間の予想通過時刻は5分単位に丸められる

3. 時間 x ICの2次元で構成される「時間指定あり検索数テーブル」を用意し、「5分単位に丸められた時刻に各区間を通過するルートを検索したログの数」をテーブル上の対応する要素に加算していく



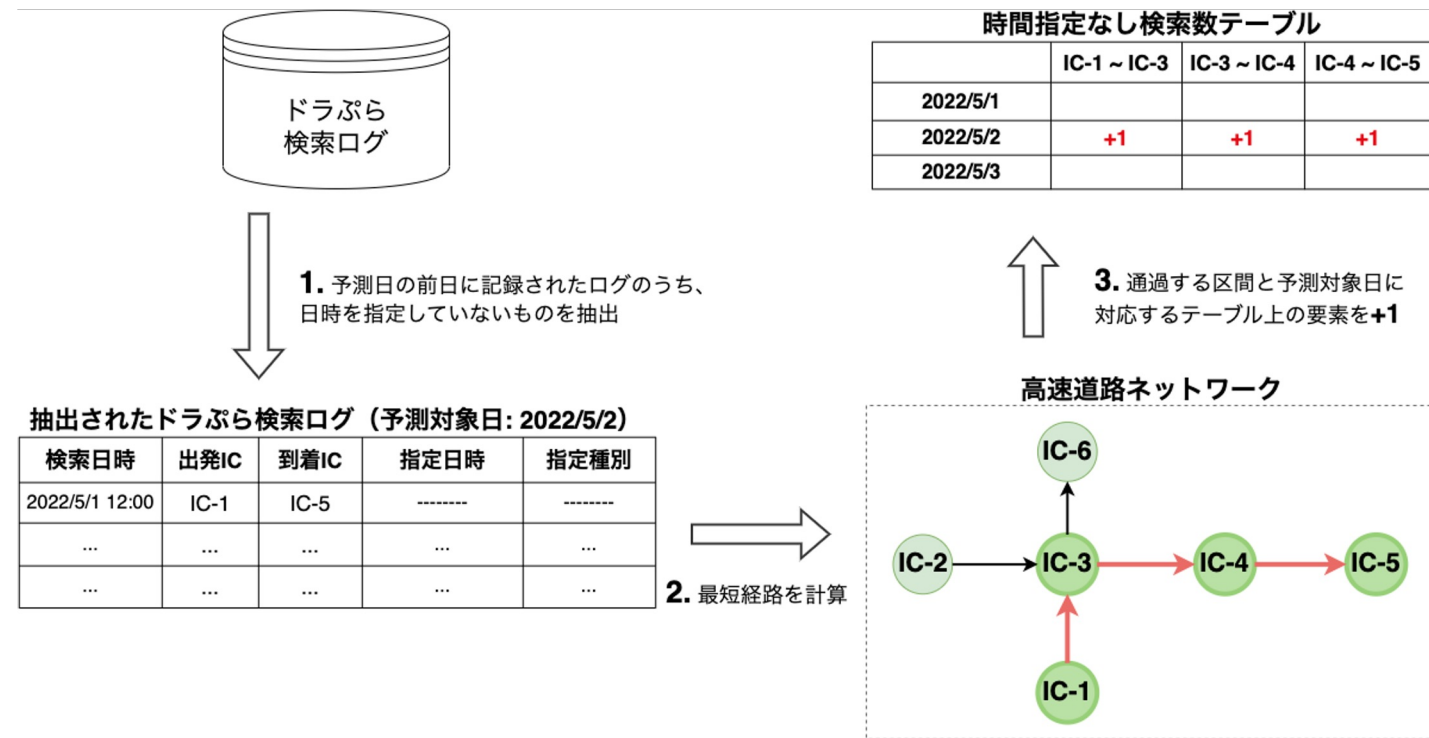
	IC-1 ~ IC-3	IC-3 ~ IC-4	IC-4 ~ IC-5
2022/5/2 10:00	+1		
2022/5/2 10:05			
2022/5/2 10:10			
2022/5/2 10:15		+1	
2022/5/2 10:20			
2022/5/2 10:25			+1



4.4 ドラぷら検索ログ構造化手順: 時間指定なし検索ログ

- 時間指定なし検索ログを以下の手順で時間指定なし検索数テーブルの形に構造化する

1. 予測対象日の前日に記録されたドラぷら検索ログのうち、日時を指定していないログのみを抽出
2. 抽出されたログの出発ICから到着ICに至る最短経路を計算
3. 時間 x ICの2次元で構成される「時間指定なし検索数テーブル」を用意し、
「各区間を通過するルートを検索したログの数」を対応する区間ごとに予測対象日（= 検索実行日の翌日）の行に加算していく
 - ▶ 日時を指定せずに検索されたルートは将来的な移動のタイミングが不明であるため、時間指定あり検索よりも粒度を粗くする
 - ▶ そのため、各区間について、「その区間を通過するルートを検索したログ数」を1日ごとに集計している



4.5 提供データに含まれるドラぷらルート検索データについて

- 提供データに含まれるドラぷらルート検索データは以下の2種類のデータから構成される
 1. 時間指定あり検索数
 - ◆ 前述の構造化の結果得られた時間指定あり検索数テーブルの値を（時間方向に）1時間単位でリサンプリングした値
 2. 時間指定なし
 3. 検索数
 - ◆ 前述の構造化の結果得られた時間指定なし検索数テーブルの値
- データフォーマット（主キー）
 1. 時間指定あり検索数
 - ◆ 時刻（1時間単位）：**datetime (str)**
 - ◆ 道路区間の始点コード：**start_code (int)**
 - ◆ 道路区間の終点コード：**end_code (int)**
 2. 時間指定なし検索数
 - ◆ 日付（1日単位）：**date (str)**
 - ◆ 道路区間の始点コード：**start_code (int)**
 - ◆ 道路区間の終点コード：**end_code (int)**

5. 使用データの仕様について

使用するデータファイルは、以下の3つから構成されます。

■ 1. 道路構造データ（road.csv）

- ▶ 区間の名前や長さ、場所などの属性データ。
- ▶ 主キー
 - ◆ 道路区間の始点コード `start_code`
 - ◆ 道路区間の終点コード `end_code`

■ 2. トラカンデータ（train.csvとtest.csvに分割された時系列データ。）

- ▶ 各道路区間を通過する自動車の速度、台数などを記録した時系列データ。
- ▶ 主キー
 - ◆ 日時 `datetime`
 - ◆ 道路区間の始点コード `start_code`
 - ◆ 道路区間の終点コード `end_code`
- ▶ データ期間
 - ◆ コンテスト期間に使用可能なデータ
 - ・ 学習用（train.csv）：2021年4月8日-2022年7月31日（配布します。）
 - ・ 暫定評価用：2022年8月1日-2022年9月30日（参加者には配布されません。暫定評価に使用します。）
 - ◆ 最終評価期間に使用可能なデータ
 - ・ 推論補助用：2022年10月1日-2023年3月30日（参加者には配布されません。最終評価で使用可能。）
 - ・ 最終評価用：2023年3月31日-2023年5月7日（参加者には配布されません。最終評価に使用します。）

5. 使用データの仕様について（続き）

■ 3.ドラぷらルート検索データ（時系列データ）

▶ 各道路区間の時間指定あり検索数（search_data.csv）

◆ 主キー

- 日時 `datetime`
- 道路区間の始点コード `start_code`
- 道路区間の終点コード `end_code`

▶ 各道路区間の時間指定なし検索数（search_unspec_data.csv）

◆ 主キー

- 日付 `date`
- 道路区間の始点コード `start_code`
- 道路区間の終点コード `end_code`

▶ データ期間

◆ コンテスト期間に使用可能なデータ

- 学習用（search_data.csv, search_unspec_data.csv）：2021年4月8日-2022年7月30日（配布します。）

◆ 暫定評価・最終評価期間に使用可能なデータ

- 暫定評価用：2022年8月1日-2022年9月30日（参加者には配布されません。）
- 最終評価用：2022年10月1日-2023年5月7日（参加者には配布されません。最終評価に使用可能。）

5.1 道路構造データの詳細①

■ 道路構造データは、高速道路の区間を挟むICの名称、位置や当該区間に設置されたトラフィックカウンターの位置等を示す、静的なデータであり詳細は次のとおり。

- ▶ 区間数（79区間（上り下りの双方）* 属性データ（18項目）
 - ◆ 関越道 63区間（= 32*2-1※）
※「高崎玉村スマートIC → 高崎JCT（下り区間）」は、トラフィックカウンターの設置がなく、予測の対象外としています。
 - ◆ 館山道 16区間（= 8*2）

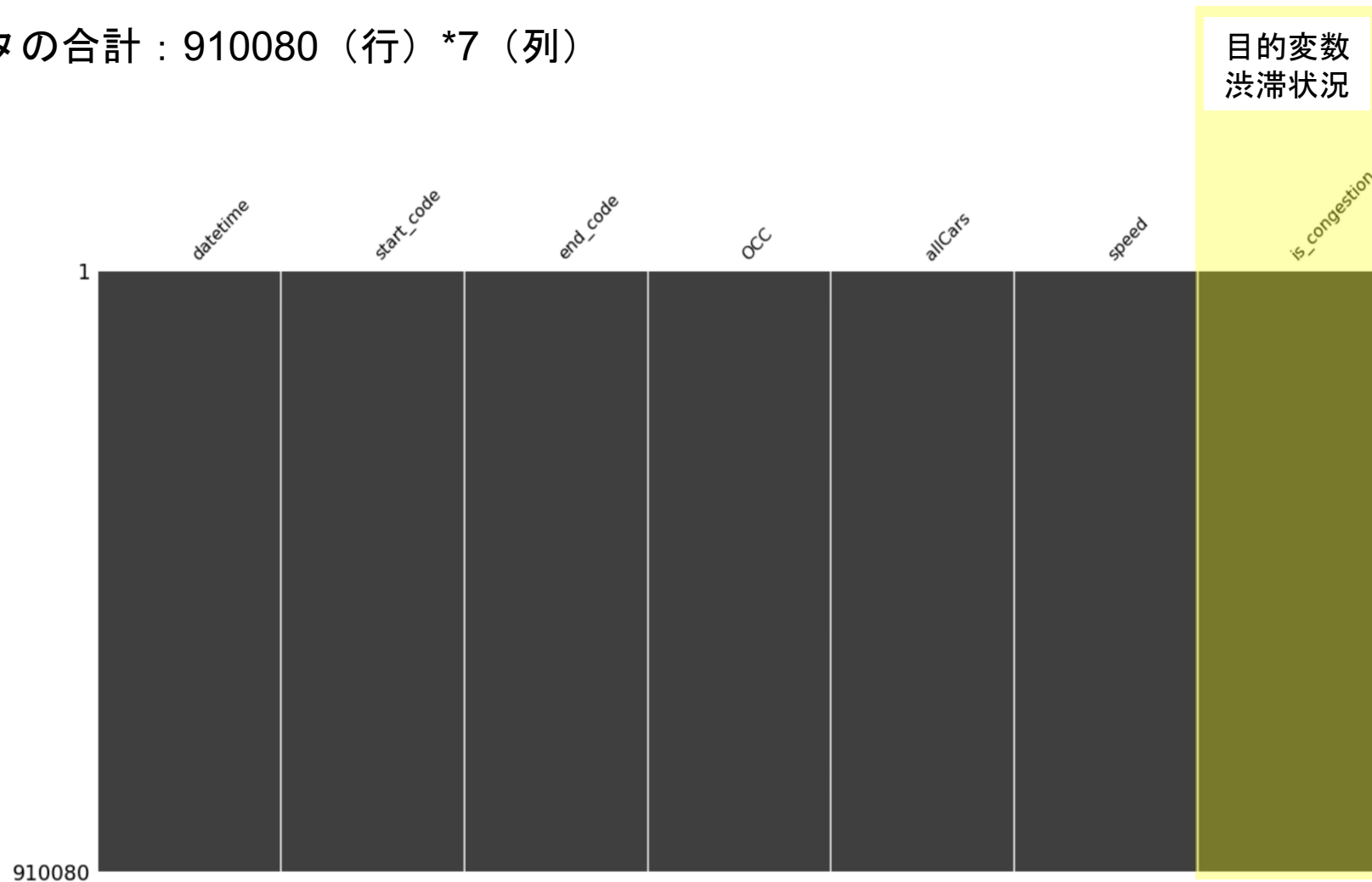
道路構造データ仕様				
ヘッダ名称	データ型	説明	値例	備考
start_name	str	区間の始点名称	所沢	
end_name	str	区間の終点名称	大泉 J C T	
start_code	int64	区間の始点コード	1800006	
end_code	int64	区間の終点コード	1110210	
start_pref_code	int64	区間の始点がおける県コード	11	https://nlftp.mlit.go.jp/ksj/gml/codelist/PrefCd.html
end_pref_code	int64	区間の終点がおける県コード	13	
start_lat	float64	区間の始点の緯度	35.80615	
end_lat	float64	区間の終点の緯度	35.75582	
start_lng	float64	区間の始点の経度	139.535511	
end_lng	float64	区間の終点の経度	139.601514	
start_degree	int64	区間の始点の度数	2	各ICが接続する他ICの数
end_degree	int64	区間の終点の度数	4	

5.1 道路構造データの詳細②

道路構造データ仕様				
ヘッダ名称	データ型	説明	値例	備考
direction	str	方向	上り;下り	上り（0kpへ向かう方向） 下り（0kpから離れる方向）
KP	float64	トラフィックカウンターのキロポスト（km）	10.21	<ul style="list-style-type: none">• KP（キロポスト）は、起点を0kpとして、道路中心線に沿う位置。• 関越道起点:「大泉JCT（練馬IC）」• 館山道は「篠崎IC（京葉道路）」を起点としているため、「京葉道路・館山自動車道接続部」のKPは35.7となる。
start_KP	float64	始点のICを代表するキロポスト（km）	43.7	
end_KP	float64	終点のICを代表するキロポスト（km）	35.7	
limit_speed	int64	制限速度（km/h）	100	館山道は均一で100、関越道は80と100が混在
road_code	int64	道路コード	1800	1800（関越道）;1130（館山道）

5.2 トラカンデータ概要

- データ粒度：1時間間隔
- 総区間数：79個
- 学習用データの合計：910080（行）*7（列）



5.2 トラカンデータ詳細

	ヘッダ名称	データ型	説明	値例	備考
時刻	datetime	str	時刻	2021-04-08 00:00:00	
区間始点と終点	start_code	int64	区間の始点コード	1800006	道路構造データを参照
	end_code	int64	区間の終点コード	1110210	
トラカンデータ	allCars	int64	1時間内の全車線の通過台数の合計	568	計測器から速度、台数、OCCが計測されている
	OCC	float64	1時間内の全車線の占有率（%）	1.6	
	speed	float64	1時間内の全車線の平均速度	87	
目的変数	is_congestion	int64	渋滞あり・なし	1/0	

5.3 ドラぷらルート検索データ概要と詳細①

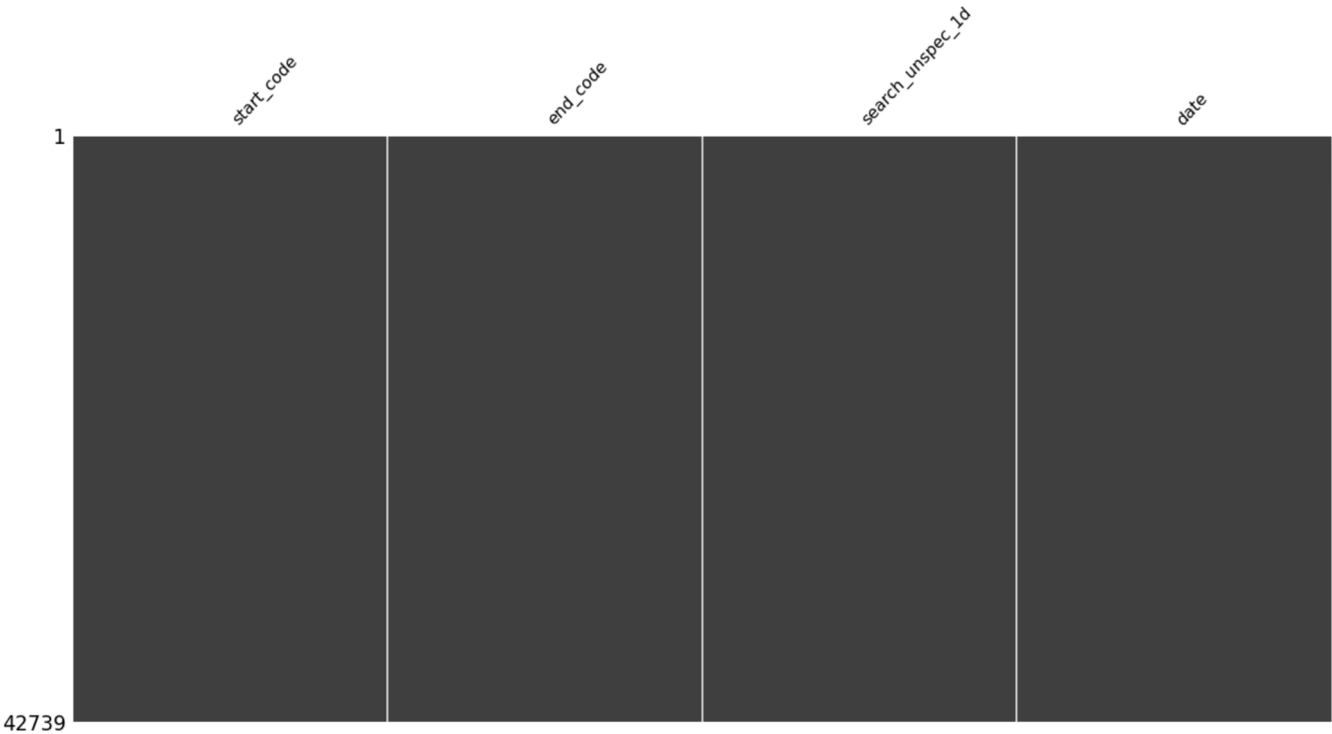
- 時間指定あり検索数（search_data.csv）
- データサイズ
 - ▶ 1025736（行）*4（列）

datetime	start_code	end_code	search_1h
1			
1025736			

ヘッダ名称	データ型	説明	値例	備考
datetime	str	時刻	2021-04-08 00:00:00	
start_code	int64	区間の始点コード	1800006	道路構造データを参照
end_code	int64	区間の終点コード	1110210	
search_1h	int64	ドラぷらルート検索データ（時間指定あり検索数（1時間ごとの集計））	4	詳細は「ドラぷら検索データの処理について」を参照

5.3 ドラぷらルート検索データ概要と詳細②

- 時間指定なし検索数（search_unspec_data.csv）
- データサイズ
 - ▶ 42739（行）*4（列）



ヘッダ名称	データ型	説明	値例	備考
date	str	日付	2021-04-08	
start_code	int64	区間の始点コード	1800006	
end_code	int64	区間の終点コード	1110210	
search_unspec_1d	int64	ドラぷらルート検索データ（時間指定なし検索数（前日分の集計））	2142	詳細は「ドラぷらルート検索データの処理について」を参照

6. 予測対象と予測タスクについて

■ 予測対象（渋滞の有無）について

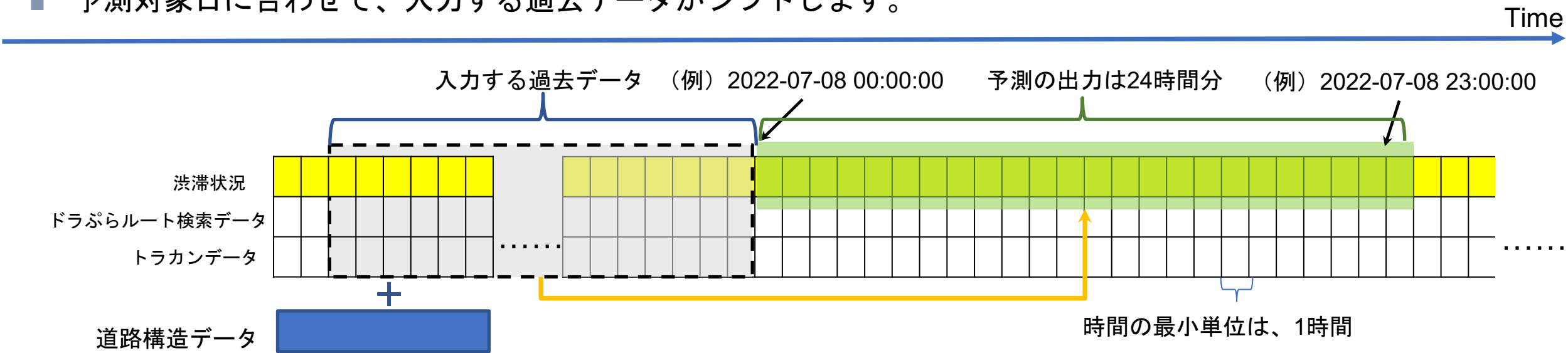
- ▶ 予測の対象とする渋滞の有無については、各区間（上り下りの両方）に設置されたトラフィックカウンターから計測された自動車の平均速度が時速40km以下の場合に、渋滞が発生している（渋滞有り）と定義します。
- ▶ なお、上記渋滞の定義は、NEXCO東日本における渋滞の定義とほぼ一致しますが、厳密な渋滞の定義とは異なり、データチャレンジのため、便宜的に用意したものです。




■ 予測タスクについて

- ▶ 過去のトラカンデータ及びドラぷらルート検索データ並びに道路構造データ（input）から、翌日の0時から24時までの高速道路（関越道・館山道）の各区間における渋滞の有無（output）を、1時間毎に予測します。
- ▶ 予測の対象範囲は、次のとおりとします。
 - ◆ 暫定評価期間：2022年8月1日～2022年9月30日の週末、祭日、連休等（下記の指定日のとおり。）
 - 2022年8月6日、2022年8月7日、2022年8月10日～2022年8月16日、2022年8月20日、2022年8月21日、2022年8月27日、2022年8月28日、2022年9月3日、2022年9月4日、2022年9月10日、2022年9月11日、2022年9月17日～2022年9月25日
 - ◆ 最終評価期間：2023年4月1日～2023年5月7日の指定日（下記の指定日のとおり。）
 - 2023年4月1日、2023年4月2日、2023年4月8日、2023年4月9日、2023年4月15日、2023年4月16日、2023年4月22日、2023年4月23日、2023年4月29日～2023年5月7日

6. 予測対象と予測タスクについて（続き）

- 予測タスクにおける入力と出力の関係は次の通り。
- 予測対象日に合わせて、入力する過去データがシフトします。



	説明
	入力する固定データ ・ 道路構造データは、固定値です。
	入力する過去データ ・ 過去データ（トラカンデータ及びドラぷらルート検索データ）の長さは任意です。
	出力データ ・ 毎日0時に予測タスクを行うことを想定しています。 ・ 0時において、次の24時間分（24個）の渋滞状況を出力し、予測対象日数に合わせて、予測を繰り返します。

7. 提出データの仕様について

- 提出データの仕様について

予測プログラムが提出物となります。

提出されたプログラムを用いて、SIGNATEサーバーにおいて予測を行い、そして評価を行います。

▶ 詳細は配布データのreadme.mdを参照してください。

8. 評価基準について

- データチャレンジの評価は、評価対象期間の全区間（上り下りの両方）における予測結果（渋滞の有無）を対象に、次の指標により行います。
- 評価指標について
 - ▶ 評価期間中の全ての渋滞（正解）と予測のF1スコアを評価指標とします。

$$F1Score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

	予測：渋滞あり	予測：渋滞なし
正解：渋滞あり	True Positive (TP) 正解は渋滞あり 予測は渋滞あり	False Negative (FN) 正解は渋滞あり 予測は渋滞なし
正解：渋滞なし	False Positive (FP) 正解は渋滞なし 予測は渋滞あり	True Negative (TN) 正解は渋滞なし 予測は渋滞なし

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$