第 53 回土木計画学研究発表会・講演集

プローブ旅行速度データの データクレンジング手法の開発

松島　敏和[1]・橋本　浩良[2]・高宮　進[3]

[1]正会員　中央復建コンサルタンツ株式会社 計画系部門 交通計画グループ（〒533-0033 大阪市東淀川区東中島4-11-10）

前 国土交通省 国土技術政策総合研究所 道路研究室（〒305-0804 茨城県つくば市旭1番地）

E-mail: matsushima_t@cfk.co.jp

[2]正会員　国土交通省 国土技術政策総合研究所 道路研究室（〒305-0804 茨城県つくば市旭1番地）

E-mail: hashimoto-h22ab@nilim.go.jp

[3]正会員　国土交通省 近畿地方整備局 兵庫国道事務所（〒650-0042　神戸市中央区波止場町3-11）

前 国土交通省 国土技術政策総合研究所 道路研究室（〒305-0804 茨城県つくば市旭1番地）

近年のICT技術の進展により，カーナビなどから取得される自動車の移動履歴情報が大量に収集されるようになっている．これらをリンク別の旅行速度に加工したプローブ旅行速度データが渋滞状況の分析などに活用されている．本稿では，旅行速度調査の主な指標値となっている平均旅行速度に着目して，平常時の一般的な道路交通状況を把握するためのデータクレンジング手法を提案し，その性能を評価する．

提案手法の特徴は，サンプルデータであるプローブ旅行速度データの特性を考慮し，同一時間帯における個別車両間の所要時間差に着目する点である．データクレンジングの際の閾値となる所要時間差を感度分析により設定し，ETC2.0プローブ情報のプローブ旅行速度データに提案手法を適用したところ，個別車両のプローブ旅行速度データの相互比較により，路側での駐停車など特異な走行と想定されるデータを除去できていることが確認できた．

Key Words: *Probe data, Travel speed, Data cleansing, ETC2.0 probe data*

1. はじめに

近年の ICT 技術の進展により，カーナビなどから取得される自動車の移動履歴情報（以下「プローブデータ」という．）が大量に収集されるようになっている．

プローブデータを DRM などのリンク別の旅行速度に加工したデータ（以下「プローブ旅行速度データ」という．）が，渋滞状況の分析など道路交通状況の把握に広く活用されている．プローブ旅行速度データには，個人の運転特性や路側での駐停車などによる特異な高速度や低速度のデータが存在するため，分析にあたっては特異なデータの除去（以下「データクレンジング」という．）が必要である．

プローブ旅行速度データに含まれる特異なデータの扱いは，現状，分析者に委ねられている．汎用的なデータクレンジング手法を確立することが，分析の効率化や成果の質の向上に資する．

本研究では，旅行速度調査の主な指標値となっている平均旅行速度に着目して，平常時の一般的な道路交通状況を把握するためのデータクレンジング手法を提案し，その性能を評価する．

2. データクレンジングの留意点

(1) プローブ旅行速度データの特異値について

プローブデータが取得される走行車両（以下「プローブカー」という．）の道路上の走行から，プローブ

データを介してプローブ旅行速度データを加工する際の手順を以下に示す.

1) 道路上の走行
2) GPS などによる測位（プローブデータの記録）
3) 移動体通信や路車間通信によるプローブデータの収集
4) プローブデータの DRM リンクなどのネットワークデータへのマップマッチング処理
5) マップマッチング処理後の測位点によるリンクごとの旅行時間（流入・流出時間の差）の算出
6) 走行距離（リンク延長）を旅行時間で除したプローブ旅行速度データの算出

次に，一連の処理過程におけるプローブ旅行速度データに特異値が含まれる要因に着目する．1) では，個人の運転特性，沿道施設への立ち寄りや路側での荷物の積み下ろしなどによる路側での駐停車など，4) では，マップマッチング処理のエラーなどが特異が含まれる要因として挙げられる．本研究では，1) ～5) の過程を経て，6) の作業段階におけるデータクレンジング手法を提案する．

なお，上記以外の要因として路上工事や事故などの特異事象の影響が考えられるが，交通規制の区間と期間は VICS 規制情報などから把握可能で，当該区間・期間を予め処理対象から除くことができるため，議論の対象外とする．

プローブ旅行速度データによる道路の混雑状況の把握には，通常，同一時間帯の平均旅行速度が用いられる．これは，1 時間や 15 分などの一定の時間幅おいて，プローブカーの旅行速度の調和平均をリンクごとに算出するものである．

平均旅行速度
＝個別車両の旅行速度の調和平均
＝リンク延長／個別車両の旅行時間の算術平均　　(1)

算術平均は特異値の影響を受けやすく，含まれる特異値の件数は僅かであったとしても，平均旅行速度に及ぼす影響は大きい．特異値が存在するデータを用いた分析結果では，道路交通状況の評価を見誤ることにつながるため，特異値とみなせるデータは，除去などの対応を講じる必要がある．

プローブ旅行速度データの特異値の事例を示す．**図-1** は，つくば市内の DRM リンクで，大規模商業施設（イーアスつくば）の駐車場入り口が存在しており，休日は駐車場入り口渋滞が頻繁に発生する区間である．**図-2** は，当該リンクの 2015 年 10 月 4 日（日）における民間プローブデータの個別車両の旅行速度と 1 時間ごとの平均旅行速度を示したものである．

平均旅行速度は，日中 15km/h から 25km/h 程度で推移しているものの，17 時台だけは 5km/h 程度となっている．これは，同時間帯に 1km/h 程度の走行車両が存在していることに起因する．同じ 17 時台には，当該リンクを 35km/h 程度で走行している車両も存在するため，1km/h 程度の走行のデータは路側での駐停車（大規模商業施設の駐車場入り口待ち渋滞が想定される）による特異値と考えられる．このように，たとえ 1 件であっても特異値と考えられるデータが含まれると，道路交通状況を適切に評価できなくなる可能性がある．

図-1　DRMリンクの位置図（つくば市）

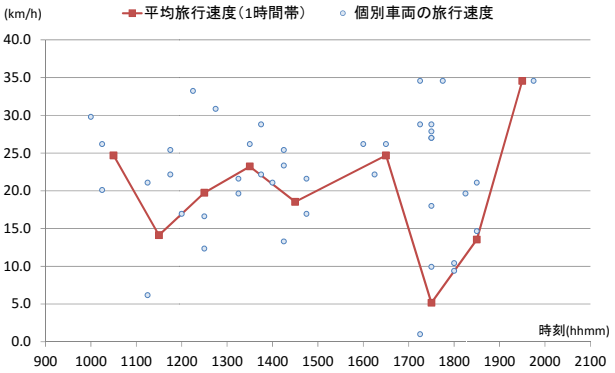

図-2　個別車両の旅行速度と平均旅行速度
（民間プローブデータ，2015年10月4日（日））

(2)　プローブ旅行速度データの旅行速度分布

図-3，**図-4** はそれぞれ高速道路と直轄国道における ETC2.0 プローブ情報の DRM リンク別，小型車 1 台毎の旅行速度の分布である（2015 年 3 月～5 月，平日 7～8 時台，全国）．これは，プローブ旅行速度データの旅行速度分布をマクロに把握するために，ETC2.0 プローブ情報のプローブ旅行速度データが取得されたすべての DRM リンクにおけるすべてのプローブカーの旅行速度を集計したものである．このため，同じ区間を走行する同一のプローブカーが含まれうる．

高速道路における旅行速度分布は，ほぼ正規分布で 90km/h 付近にピークがある．直轄国道における旅行速

度分布は，二山の分布で 10km/h 付近と 50km/h 付近にピークがある．直轄国道では，自由流に近い領域と低速度の領域それぞれの分布を合成したものと解釈可能で，信号待ちの速度低下が影響していると考えられる．

旅行速度分布の特性を踏まえると，データクレンジングに際しては下記の留意点が挙げられる．

- 一般に旅行速度の分布形状は区間・時間によって異なるため，標準偏差などの統計値を一律の閾値としてサンプルデータに対してデータクレンジングを実施することは，基準の解釈が曖昧となり適切ではない．
- 道路交通状況は一定ではなく，常に変化しうるため，プローブ旅行速度データの特異値の領域を一意に設定することは難しく，同一時間帯の道路交通状況を考慮する必要がある．

既往研究 [1)2)]では，ナンバープレート調査により全数に近いデータが取得された場合に，路側での駐停車などによる旅行時間の外れ値（異常な低速度のデータ）を確実に除外するための統計的な閾値について感度分析が実施されている．一方，本研究で対象にするプローブ旅行速度データは分布が不明の母集団から得られるサンプルデータであるため，このような方法は適用しづらい．そこで本研究では，同一時間帯に取得されたプローブ旅行速度データの相互比較により，異常な低速度のデータを除去することを提案する．

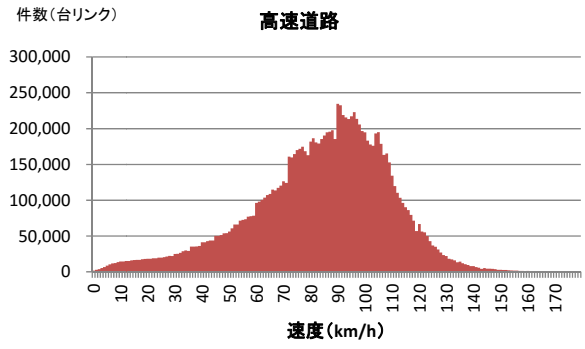

図-3　高速道路における旅行速度分布

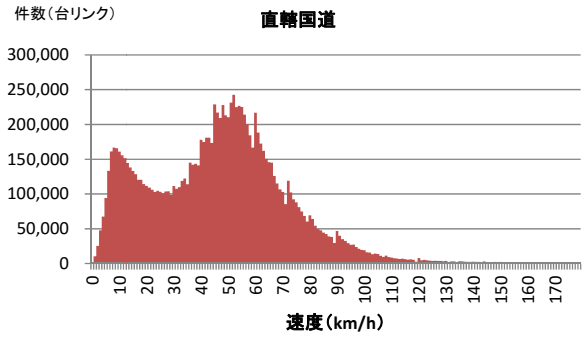

図-4　直轄国道における旅行速度分布

3.　データクレンジング手法の提案と感度分析

(1)　データクレンジング手法の提案

上記の留意点を考慮して，①閾値による高速度・低速度データの除去，②同一時刻帯の道路交通状況を考慮した低速度データの除去の 2 段階のデータクレンジング手法を提案する（**表-1**）．

②が本手法の特徴である．**図-5** の例のような駐停車のデータは道路交通状況分析には適さないため，データの除去対象とすべきであるものの，個別車両の旅行速度のみでは混雑による低速度走行のデータと区別ができない．そこで，駐停車車両は他の車両との所要時間差が大きく，混雑による低速度車両は他の車両との所要時間差が小さいことに着目し，当該リンクの所要時間が同一時間帯における最小所要時間（最高速度）のサンプルから一定時間以上遅れる車両を特異な走行とみなし，該当するデータを除去することを考えた．

リンク延長が長くなると最小所要時間からの時間差が大きくなるため，最小所要時間が非常に小さい場合，通常の走行であってもデータクレンジング対象となる可能性がある．この影響を抑制するため，高速道路では 80km/h，一般道では 30km/h の走行時の所要時間を最小所要時間の下限とする．

表-1　データクレンジング手法の概要

項目	内容
①閾値によるデータの除去	• 低速度側 1km/h 未満，高速度側 150km/h 以上となるデータを除去
②同一時間帯の道路交通状況を考慮したデータの除去	• 同一時間帯における最小所要時間（最高速度）のサンプルから一定時間以上遅れるサンプルのデータを除去 • ただし，高速道路では 80km/h，一般道では 30km/h の走行時の所要時間を最小所要時間の下限とする

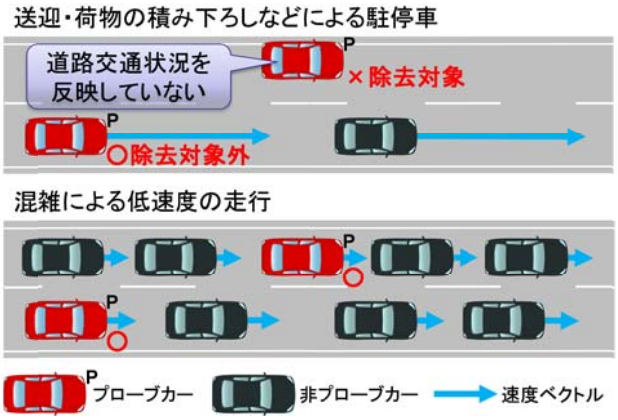

図-5　除去対象とする走行データの例

(2) 時間閾値の感度分析

上記②のパラメータである同一時間帯最小所要時間からの差の閾値（以下「時間閾値」という．）のあたりをつけるために，データクレンジングを試行する．時間閾値を変化させてデータクレンジングを施した前後のデータ件数の変化，除去対象となったデータの平均旅行速度および旅行速度の中央値の変化に着目した感度分析を実施する．

使用するデータは ETC2.0 プローブ情報（2015 年 4 月～6 月），対象地域は宮城県，対象路線は高速道路と直轄国道とした．時間閾値は，200 秒，400 秒，600 秒，800 秒，1,000 秒で変化させるものとする．データ件数は，DRM リンク別上下方向別で高速道路 151.4 万台リンク，直轄国道 114.0 万台リンクである．

図-6，**図-7** は時間閾値を変化させてデータクレンジングを実施したときの，データ除去件数と除去前に対する除去後のデータ割合である．いずれも時間閾値が大きくなるほど，除去されるデータ件数が減少し，600 秒以上では除去前に対する除去後のデータ割合の変化が小さくなることが確認できる．

図-8，**図-9** は，DRM リンク延長別除去対象データの平均旅行速度である．いずれも時間閾値が小さいほど除去対象となる車両の速度が高く，時間閾値が大きくなるほど速度が低くなる．また，リンク延長が長くなると，比較的速度の高い車両が除去対象に含まれていることがわかる．時間閾値が 600 秒以上になると，どの延長においても除去対象データの平均旅行速度の変化が小さくなる傾向にある．

図-10，**図-11** は，DRM リンク延長別除去対象データの旅行速度の中央値である．旅行速度の中央値は，平均旅行速度と比較してやや速度が高いものの，いずれの道路種別でも時間閾値が 600 秒以上になると，平均旅行速度同様に中央値の変化が小さくなる傾向にある．

これらから，時間閾値が 600 秒以上では，除去対象のサンプル自体に大きな変化がなく，平均旅行速度が十分に低下し，駐停車はしていないと考えられる低速走行の車両が除去対象から外れている様子がわかる．

時間閾値は分析目的に応じて（除去対象としたい交通行動に応じて）可変である．汎用性を考慮した時間閾値のデフォルト値としては，600 秒程度が妥当であると考えられる．

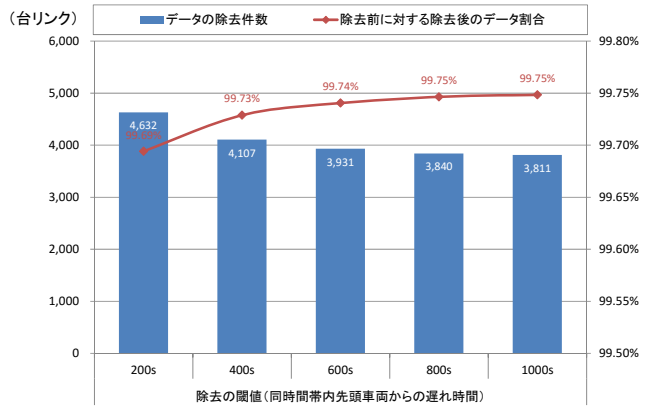

図-6 データ除去件数と除去前に対する除去後のデータの割合（高速道路）

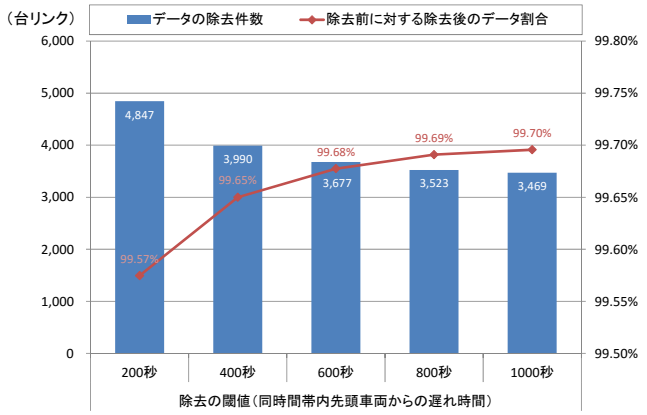

図-7 データ除去件数と除去前に対する除去後のデータの割合（直轄国道）

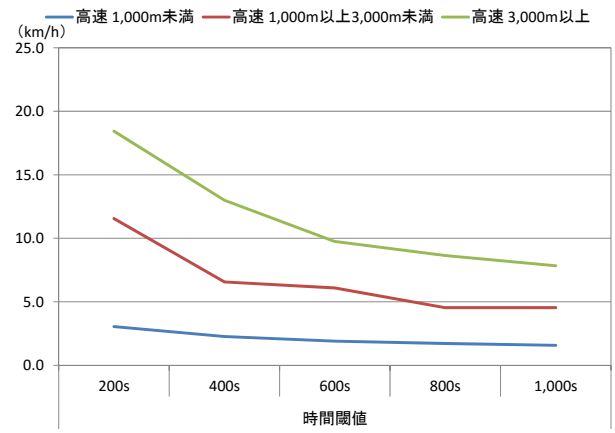

図-8 除去対象データの平均旅行速度（高速道路）

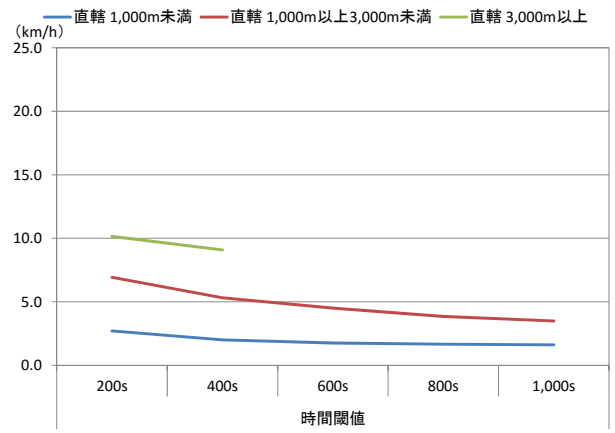

図-9 除去対象データの平均旅行速度（直轄国道）

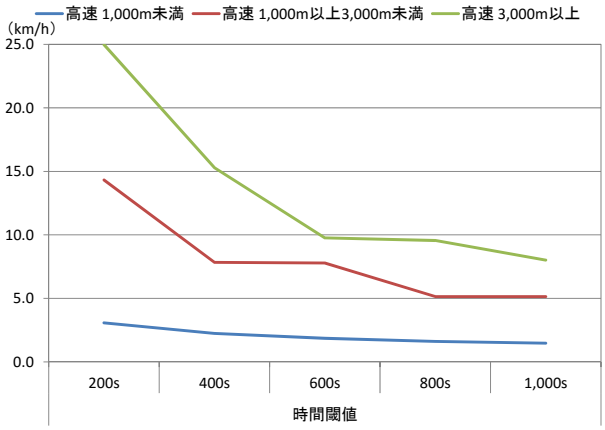

図-10　除去対象データの旅行速度の中央値（高速道路）

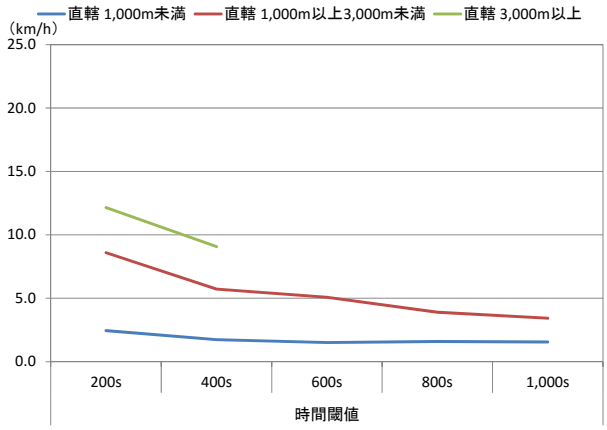

図-11　除去対象データの旅行速度の中央値（直轄国道）

(3)　データクレンジングの対象となる領域

時間閾値を600秒と設定した際の，データクレンジングの対象となるリンク延長とそのリンクを通過する際の所要時間の領域を確認する．

図-12は，一般道におけるデータクレンジングの対象となる領域を示したものである．赤色の領域は，一律の閾値による除去対象（**表-1**の①に該当）であり，青色の領域は，最小所要時間と考える30km/h相当の所要時間を600秒以上上回るもの（**表-1**の②に該当）である．

また，緑色の領域は，最小所要時間が30km/h相当を下回る場合の，有効データの領域である．たとえば，リンク延長が2kmで最小所要時間が20km/h相当であった場合，所要時間に600秒加えた960秒（7.5㎞/h）までのプローブ旅行速度データを有効データとして扱う．

4.　データクレンジングの試行と結果の考察

表-2 は，ETC2.0 プローブ情報（2015 年 4 月～6 月，全国，全道路種別，昼間 12 時間）の速度ランク別 DRM リンク延長別旅行速度データのデータ件数で，合計約 274 百万台リンクである．DRM リンク延長でみると 500m 未満が圧倒的に多く，速度ランクでみると 40km/h 以上 60km/h 未満が最も多い．また，DRM リンク延長が 2,000m 以上の場合，1km/h 未満のデータは存在しない．

表-3 は，このプローブ旅行速度データに，時間閾値を 600 秒と設定してデータクレンジングを施した際の，速度ランク別 DRM リンク延長別の削除対象件数の割合である．全体の除去対象件数の割合は，0.30％（約 0.8 百万台リンク）である．前述のとおり，割合的には 0.30％と小さいものの，分析に際して，その影響は大きいものと想定される．

表-3 で着目すべき点は，同一時間内における走行車両の旅行速度の相互比較によるデータクレンジング（**表-1** の②）の効果で，1km/h 以上の走行車両のデータであっても除去対象となっている点である．DRM リンク延長が長いほど，速度が低いほど，除去対象件数の割合が大きくなる．たとえば，DRM リンク延長が 3,000m 以上で，旅行速度が 1km/h 以上 10km/h 未満のデータは半数以上が特異値と判別される．

上記の結果より，同一時間帯における走行車両の旅行速度を相互比較することで，路側での駐停車など特異な走行と想定されるデータを除去できていることが確認できた．

5.　おわりに

本研究では，プローブ旅行速度データに含まれる特異なデータを除去するデータクレンジング手法を提案した．同一時間帯における走行車両の旅行速度の相互比較により，データクレンジングの精度を向上させることができることを確認した．

ここでの開発手法の大きな利点は，リンクごとの旅行速度データの母集団の分布が不明であっても，同一時間帯に複数のプローブ旅行速度データが取得されれば適用可能な点である．複雑なアルゴリズムによらず，特異なプローブ旅行速度データの除去性能を有するため，汎用性は高いと考える．

今後の課題としては，分析目的に合致する時間閾値の整理，リンク延長と最小所要時間に相当する速度との関係，暫定 2 車線区間などの道路状況とデータクレンジング性能などについて，さらなる分析が必要と考える．

プローブデータの取得状況は大きく増加傾向であり，今後さらに豊富なデータを用いた分析が可能になる．ひきつづき，手法の吟味を進めるとともに，マニュアル化などの汎用化策を検討する．本研究が，分析者がプローブ旅行速度データを取り扱う際の一助となればと思う．

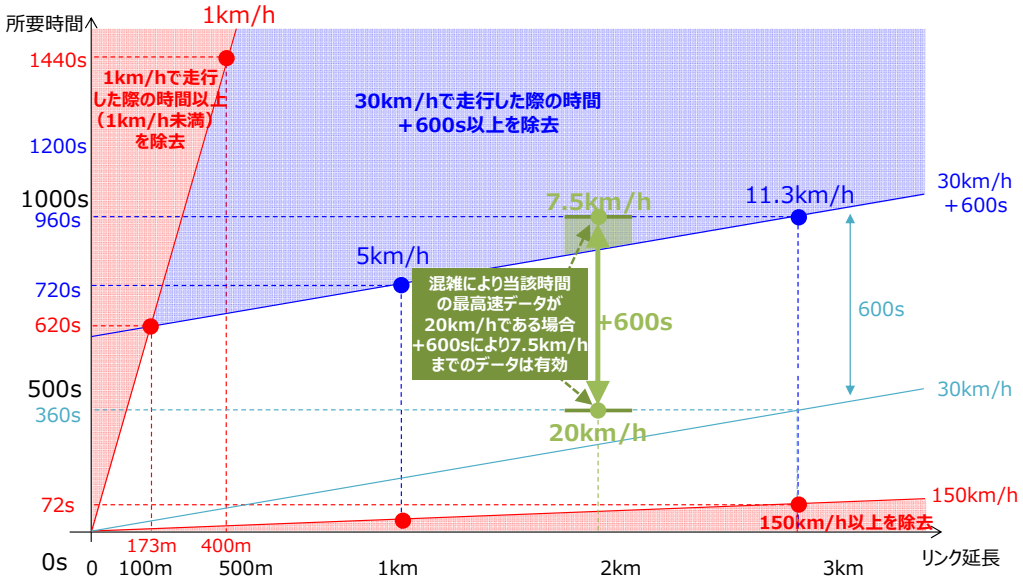

図-12　データクレンジングにより除去対象となる領域（一般道の例）

表-2　データクレンジング前のデータ件数

速度ランク＼DRMリンク延長	500m未満	500m以上 1000m未満	1000m以上 2000m未満	2000m以上 3000m未満	3000m以上	合計
1km/h未満	101,128	1,730	28	0	0	102,886
1km/h以上10km/h未満	11,992,068	381,337	61,583	7,295	1,913	12,444,196
10km/h以上20km/h未満	21,784,312	1,393,434	182,147	21,286	8,499	23,389,678
20km/h以上40km/h未満	39,880,142	5,829,960	1,169,011	105,362	47,875	47,032,350
40km/h以上60km/h未満	63,059,274	8,611,434	2,945,802	497,285	206,778	75,320,573
60km/h以上80km/h未満	36,702,063	7,951,405	4,152,835	1,162,003	861,972	50,830,278
80km/h以上100km/h未満	23,351,446	7,013,708	5,187,359	2,015,993	1,852,373	39,420,879
100km/h以上120km/h未満	10,918,100	3,441,007	2,727,682	1,246,404	1,332,974	19,666,167
120km/h以上150km/h未満	3,262,987	600,510	427,468	180,034	189,023	4,660,022
150km/h以上	601,744	44,377	17,343	5,410	4,609	673,483
合計	211,653,264	35,268,902	16,871,258	5,241,072	4,506,016	273,540,512

（台リンク）

表-3　データクレンジングの除去対象件数の割合

速度ランク＼DRMリンク延長	500m未満	500m以上 1000m未満	1000m以上 2000m未満	2000m以上 3000m未満	3000m以上	合計
1km/h未満	100.00%	100.00%	100.00%	－	－	100.00%
1km/h以上10km/h未満	0.14%	5.36%	15.15%	43.88%	55.15%	0.40%
10km/h以上20km/h未満	0.00%	0.00%	0.00%	3.77%	23.77%	0.01%
20km/h以上40km/h未満	0.00%	0.00%	0.00%	0.00%	0.39%	0.00%
40km/h以上60km/h未満	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
60km/h以上80km/h未満	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
80km/h以上100km/h未満	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
100km/h以上120km/h未満	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
120km/h以上150km/h未満	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
150km/h以上	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
合計	0.34%	0.19%	0.16%	0.18%	0.17%	0.30%

大　　小

凡例：割合

※除去対象件数の割合（%）＝ 除去対象データ対象件数 ／ データクレンジング前のデータ件数

参考文献

1) 橋本浩良：旅行時間データのサンプル数と旅行時間算定値の代表性の関係，土木技術資料 56-10，2014
2) 野間真俊，奥谷正，橋本浩良：道路ネットワークの評価における時間信頼性指標の適用に関する研究，土木計画学研究・講演集，Vol.37，2008
3) 松島敏和，橋本浩良，高宮進：プローブ旅行速度データのクレンジング手法に関する一考察，第 13 回 ITS シンポジウム 2015，対話セッション 2-1A-06，2015

DEVELOPMENT OF A DATA CLEANSING METHOD FOR PROBE TRAVEL SPEED DATA

Toshikazu MATSUSHIMA, Hiroyoshi HASHIMOTO and Susumu TAKAMIYA