クラスタリングによる
広域・継続的プローブカーデータの中間集計法

日下部　貴彦[1]・朝倉　康夫[2]

[1]正会員　東京大学講師　空間情報科学研究センター（〒277-8568　千葉県柏市柏の葉5-1-5）
E-mail: t.kusakabe@csis.u-tokyo.ac.jp

[2]正会員　東京工業大学教授　環境・社会理工学院（〒152-8552 東京都目黒区大岡山2-12-1-M1-20）
E-mail: asakura@plan.cv.titech.ac.jp

近年，商用車プローブカーデータをはじめとした広域かつ長期間・継続的に収集された移動体データがみられるようになった．このようなデータは，トラカンなどの既往の定点での観測装置で収集されるデータと比較すると，広域的・継続的なデータが収集されていることが特徴である一方で，データサイズが大きくまた交通運用を意図して収集したデータではないという問題がある．したがって，広域的・継続的なデータという特徴を活かしつつ交通の変動を捉えるためには，大量のデータを効率的に集計し，変動を検出するための手法が必要である．本研究では，全国で収集された数TB規模の大規模なプローブカーデータに対しクラスタ分析を行うためのデータ解析手順を提案し，実データを用いた解析を試行した．

Key Words : *Probe Vehicle，OLAP，Vector Quantization, k-means++*

1. はじめに

従来，自動車交通の観測は，車両感知器をはじめとした定点での観測方法が主要であったが，定点での観測方法では，道路ネットワーク全体に及ぶ観測装置の設置は現実的ではないことから，空間的な観測範囲という視点では限定的となっていた．1990年代後半には，Global Positioning System (GPS)の登場および情報通信技術の進展により，個々の自動車の軌跡を観測するプローブカーデータ[1][2]の取得が容易になった．初期のプローブカーによるデータ収集では，実験的なデータの収集・蓄積にとどまっており，データ収集の期間や空間的な範囲が限定的[3][4][5]であったが，近年では，事業車の運行管理システム[6]やオンラインでのナビゲーションシステム[7]が普及したことから，長期間・継続的でかつ広い空間的な範囲でのデータ取得が可能となりつつある．移動体観測では，そのようなシステムを搭載した自動車が走行する範囲内のデータが収集できることから，定点観測よりも広域のネットワーク上の情報を収集できると期待できる．

プローブカーデータでは，個々の車両の一定間隔（例えば，本研究で用いたデータでは1秒毎）の位置情報を収集していることから，定点データと比較してデータサイズは膨大となる．このような膨大なデータには，定点観測装置が設置されていない地点の情報が含まれるなど有益な情報が含まれる可能性がある．一方で，より広範囲のデータの特性を分析する際には必ずしも重要な情報だとは限らず，このようなデータを直接的に集計することは，膨大な計算時間がかかることから効率的でない場合もある．特に，広範囲でかつ継続的なデータには，分析者も予期していないような変動が観測されていたり，イベントや災害等にインシデント等，あらかじめ観測範囲を限定することが難しい現象も含まれていることも期待できることから，試行錯誤的に分析時の範囲や期間，時間・空間的な粒度を決めたり，データマイニングを行い発見的な分析を適用することも想定されるが，このような分析を行うためには，数TB/年に及ぶ観測データをそのまま用いることは計算処理に要する時間をから難しくよりおおくの計算機資源を必要とすることから効率的ではない．

そこで本研究では，まず，プローブデータを時空間メッシュごとのデータとして集計した中間データを生成する方法についてまとめる．このような中間データに必要な要件は，時間・空間の範囲を再集計可能であることである．つまり，より時間的・空間的に粒度の荒い集計値（例えば，日単位や年単位）を中間データから整合的に再集計できることである．

中間データであっても，例えば，地域メッシュの三次メッシュで集計した場合には，時刻毎に数十万メッシュ分のレコードがあるデータとなり，数秒単位の瞬時に平均や分散などの集計値を全国規模で計算することは容易

ではない．そこで，本研究では，蓄積されたデータの集計の試算等を瞬時に行うOLAP (Online Analytical Processing)の要素がある分析を念頭に，多少の整合性を犠牲にしてもさらに高速に広範囲のデータを集計する方法として，データのクラスタリングの方法を用いたベクトル量子化を用いる中間データの生成と手順についてまとめる．この方法では，集計値の試算時にはクラスタ中心のベクトルを集計し，実データの分析を代用する．これにより，大量のデータをメモリ上に読み込むことなく集計値を得ることで，計算時間を削減することを意図している．

2. 方法

(1) 再集計可能なメッシュデータの生成

a) 中間データの生成

本研究では，中間データとして，メッシュ毎に再集計可能な指標値をまとめた中間データを作成する．なお，中間データでのメッシュサイズは，標準地域メッシュの第3次メッシュとし，集計時間単位はw分とする．

中間データでは，各集計時間単位ごとに各メッシュについて「各辺の境界流入交通量」，「各辺の境界流出交通量」，「発生交通量」，「集中交通量」，「総走行距離」，「総走行時間」，「車両の瞬間存在台数」を算出し，中間データとする．

mをメッシュ番号，τを時刻番号とし，矩形であるメッシュの各辺の方角を$\delta = \mathrm{N, S, E, W}$であらわすとき，各辺の境界流入交通量を$\mathrm{f}_{i\tau}^{\delta}$，各辺の境界流出交通量を$\mathrm{g}_{i\tau}^{\delta}$とする．これらの値は，各時刻にメッシュの境界の辺をまたいだ車両数を集計することにより求める．発生交通量，集中交通量は，$x_{\mathrm{i\tau}}$，$\mathrm{y_{i\tau}}$であらわす．これらの集計値は，別途移動滞在判別や車両のエンジンのON/OFF等によりデータの仕様としてトリップの分割点と判別された点を集計して求める．総走行距離$\mathrm{d_{i\tau}}$は，メッシュ内を集計単位時間内に走行した車両の走行距離を合計することにより求め，総走行時間$\mathrm{t_{i\tau}}$はそれらの車両の走行時間を合計することにより求める．

b) 中間データからの指標算出

中間データの集計単位の領域を最小単位として任意の領域A，集計時間数Tで，流入量，流出量，発生交通量，集中交通量，プローブカー平均交通量，平均旅行速度を求める方法について述べる．

時刻番号τの流入量は，領域aの境界の各辺の集合をe(A)とするとき，

$$F_{A,t,t+T} = \sum_{(i,\delta)\in e(A),\tau\le u<\tau+T} \mathrm{f}_{iu}^{\delta} \quad (1)$$

とあらわすことができる．また，流出量も同様に，

$$G_{A,t,t+T} = \sum_{(i,\delta)\in e(A),\tau\le u<\tau+T} \mathrm{g}_{iu}^{\delta} \quad (2)$$

となる．発生交通量は，

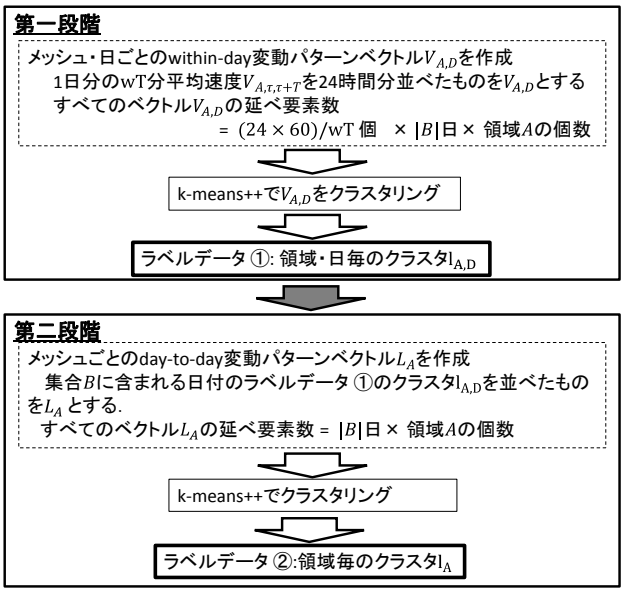

図-1 クラスタリングによる中間データの生成の流れ

$$X_{A,t,t+T} = \sum_{i\in A,\tau\le u<\tau+T} x_{iu} \quad (3)$$

であり，集中交通量は，

$$Y_{A,\tau,\tau+T} = \sum_{i\in A,\tau\le u<\tau+T} y_{iu} \quad (4)$$

である．プローブカーエリア交通量は，Edie (1963)の定義を用いると，

$$Q_{A,\tau,\tau+T} = \frac{\sum_{i\in A,\tau\le u<\tau+T} d_{iu}}{wT|A|} \quad (5)$$

として算出できる．ただし，$|A|$は領域Aの道路延長である．プローブカーエリア密度は，

$$K_{A,\tau,\tau+T} = \frac{\sum_{i\in A,\tau\le u<\tau+T} t_{iu}}{wT|A|} \quad (6)$$

であり，エリア平均速度は，

$$V_{A,\tau,\tau+T} = \frac{\sum_{i\in A,\tau\le u<\tau+T} d_{iu}}{w\sum_{i\in A,\tau\le u<\tau+T} t_{iu}} \quad (7)$$

である．

(2) クラスタリングによる中間データの生成

本研究で用いるベクトル量子化手法は，クラスタリング手法のひとつであるk-means++法[8]である．この方法は，従来のk-means法の収束性とクラスタ中心（クラスタ内分散を最小化する点）の近似比を改善した手法であり，大量のデータを処理するのに優れていることから採用した．

中間データの生成の流れ（**図-1**）は，まず第一段階で，各メッシュの各日をその日の変動パターンに基づくクラスにk-means++法を用いて分類し，その分類に基づくラベルデータ①を作成する．第二段階では，メッシュ毎にラベル①の日々の変動パターンに基づくクラスにk-means++法を用いて分類し，その分類に基づくラベルデ

ータ②を作成する．集計分析を行う段階では，ラベル①，②について各ラベルのクラスタ中心のベクトルを代用して集計することで集計時の計算時間を短縮する．

第一段階のラベルデータ①の作成手順を詳しく述べる．領域Aの第D日のWithin-day変動パターンのベクトルを

$$V_{A,D} = \{V_{A,\tau,\tau+T} | \tau \in a(D,T)\} \quad (8)$$

として作成する．ただし，$a(D,T)$は，D日に含まれる時刻番号のうちT間隔で得られる時刻番号の集合である．すべての領域と日を対象にk-means++法を適用し，k_1個のクラスタに分類し，各変動パターンベクトルにクラスタ番号($l_1 = 1,2,\ldots,k_1$)のラベルをつける．なお，クラスタ数k_1は所与とする必要がある．これにより，各領域と日についてクラスタ番号を付した$l_{A,D}$を要素としたラベルデータ①を得る．また，各クラスタ中心のベクトルV_{l_1}を得る．

第二段階のラベルデータ②の作成では，領域AのDay-to-day変動パターンベクトルを

$$L_A = \{l_{A,D} | D \in B\} \quad (9)$$

とする．ただし，Bはデータ収集期間の日を示す集合である．すべての領域を対象に，k-means++法を適用し，k_2個のクラスタに分類し，各変動パターンベクトルにクラスタ番号($l_2 = 1,2,\ldots,k_2$)のラベルをつけることにより，各領域についてクラスタ番号を付したl_Aを要素とするラベルデータ②を得る．また，各クラスタ中心のベクトルL_{l_2}を得る．

集計分析を行う段階では，データを各クラスタ中心で代用して分析を行う．ラベルデータ①とそのクラスタ中心ベクトルを用いることで各領域の日ごとの速度の単純平均と分散を算出する．領域Aの第D日の速度の日単純平均と分散の算出には，クラスタ番号$l_{A,D}$のクラスタ中心のベクトル$V_{l_{AD}}$の要素の単純平均と分散を用いる．なお，Edieの定義にしたがう速度の平均値と単純平均は定義が異なるので留意されたい．さらに，ラベルデータ②とそのクラスタ中心ベクトルを用いれば，同様の手順で，データの収集期間中の速度の日単純平均の平均と分散，日分散の平均およびその分散を集計できる．領域Aのそれらの集計値は，クラスタ番号l_Aのクラスタ中心のベクトルL_{l_A}を用いて算出する．L_{l_A}の要素は，第一段階目のクラスタ番号で構成されるが，クラスタ中心ベクトルの各要素のクラスタ番号に対応する速度の日単純平均の平均と分散を算出することで速度の平均と分散を得る．同様に，対応する速度の日分散の平均単純を求めることで日分散の平均を算出する．また，その分散を求めることで日分散の分散を算出する．

以上に示した手順での集計方法では，日単位の集計については，k_1通り行えばよく，また期間を通した分析でも$k_1 \times k_2$通りの計算でよいため，データ全体を直接集計する場合と比べて，大幅な計算時間の短縮を見込むことができる．一方で，ベクトル中心の値を代表値として用いることから，実際のデータを用いて集計した場合との乖離が発生する．このことからどの程度異なる特徴が現れるかについては，検証の必要がある．また，提案手法での算出値は，このためOLAPでの分析手順のための試算値として用い，定量分析にはその試算値を参考にして決定された範囲に基づいて，実際のデータで指標を算出する必要がある．

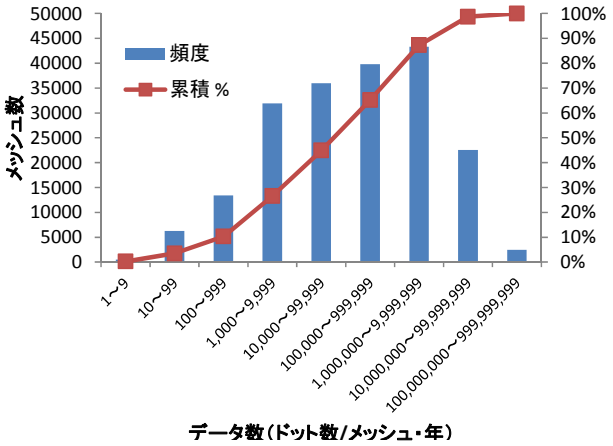

図-2　各メッシュに含まれるドット数の分布

3.　適用例

(1)　データの概要

本研究で用いたデータは，（株）富士通交通・道路データサービスが収集したプローブカーデータである．プローブデータは，約4万7千台（2014年10月時点）の商用車に設置されたGPS搭載のデジタルタコグラフから収集されたものであり，各車両の1秒毎の緯度・経度の情報である．ただし，ここで用いたデータは，停止時間に基づく起終点判別を行い，秘匿処理のためにそれらの起点・終点の直近を省いたデータセットとなっている．対象地域は，日本全域であり，使用したデータの収集期間は，2014/1/1～12/31の1年間である．この間のトリップ数は，延べ70,336,364トリップであり，一日あたりにすると327,923トリップ/日となった．ドット数は，133,574,067,086ドット観測されていた．

(2)　メッシュ化

a)　メッシュ化の実行

2章1節で述べた方法により，メッシュデータを作成した．この際，メッシュの最小単位は，標準地域メッシュの第3次メッシュとし，集計時間単位は$w = 5$分とした．メッシュ処理の結果，1ドット以上の観測があったメッシュ数は，196,209メッシュであった．第3次メッシュのメッシュあたりの面積は約1km^{2}であることから，日本の

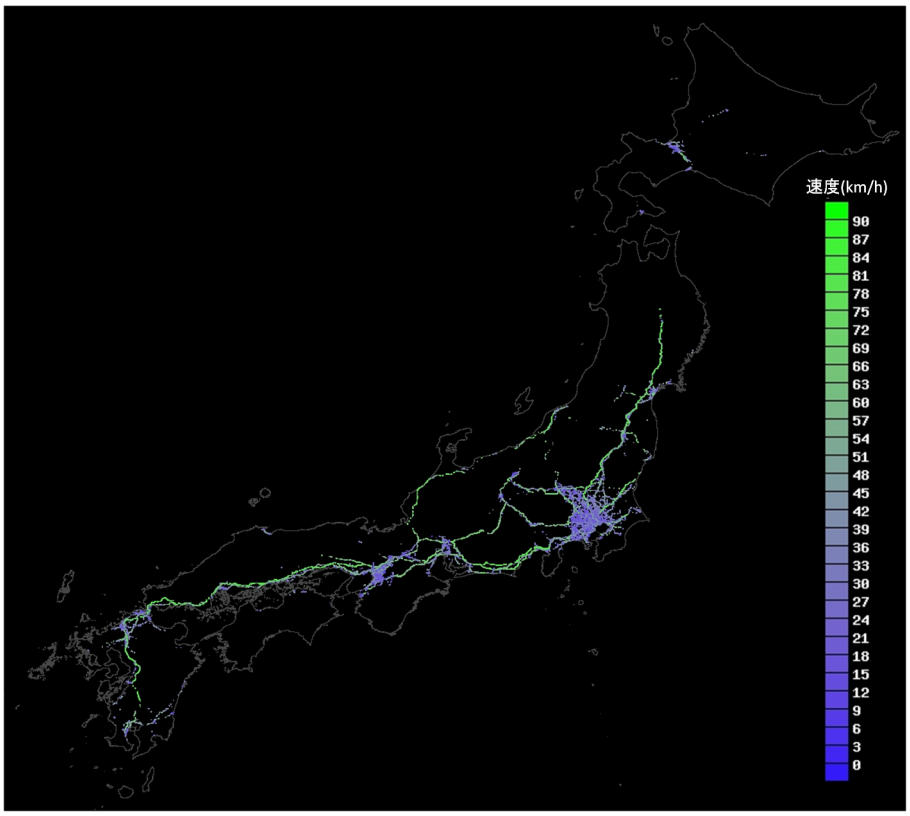

図-3 2014/2/13（平日）の第三次メッシュの日平均速度

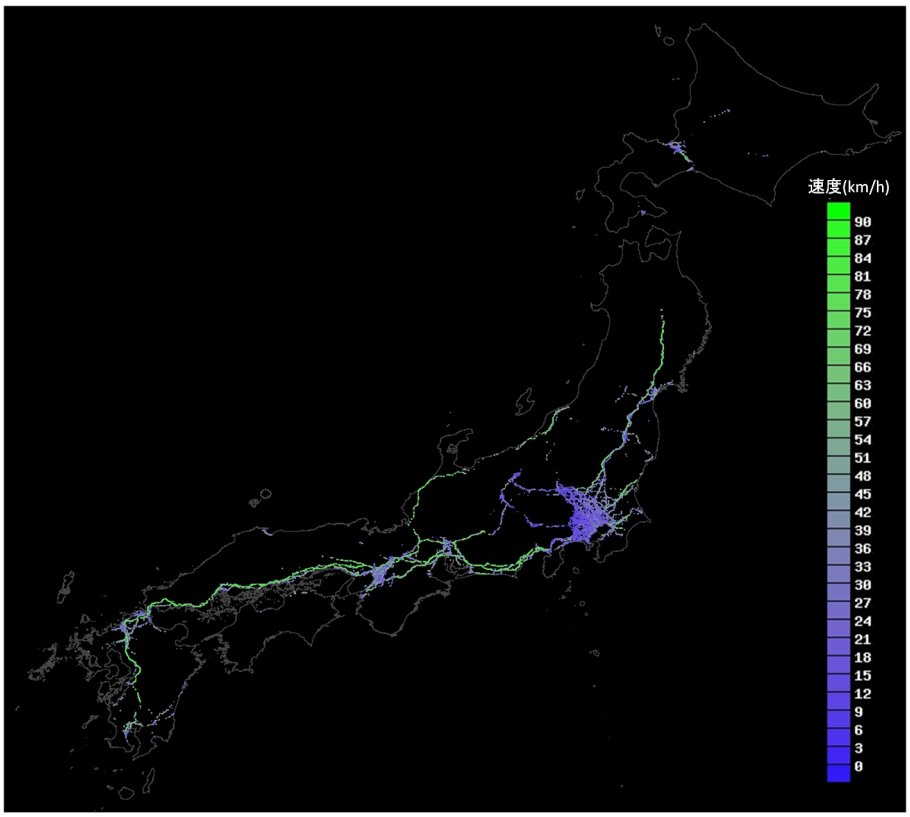

図-4 2014/2/14（豪雪）の第三次メッシュの日平均速度

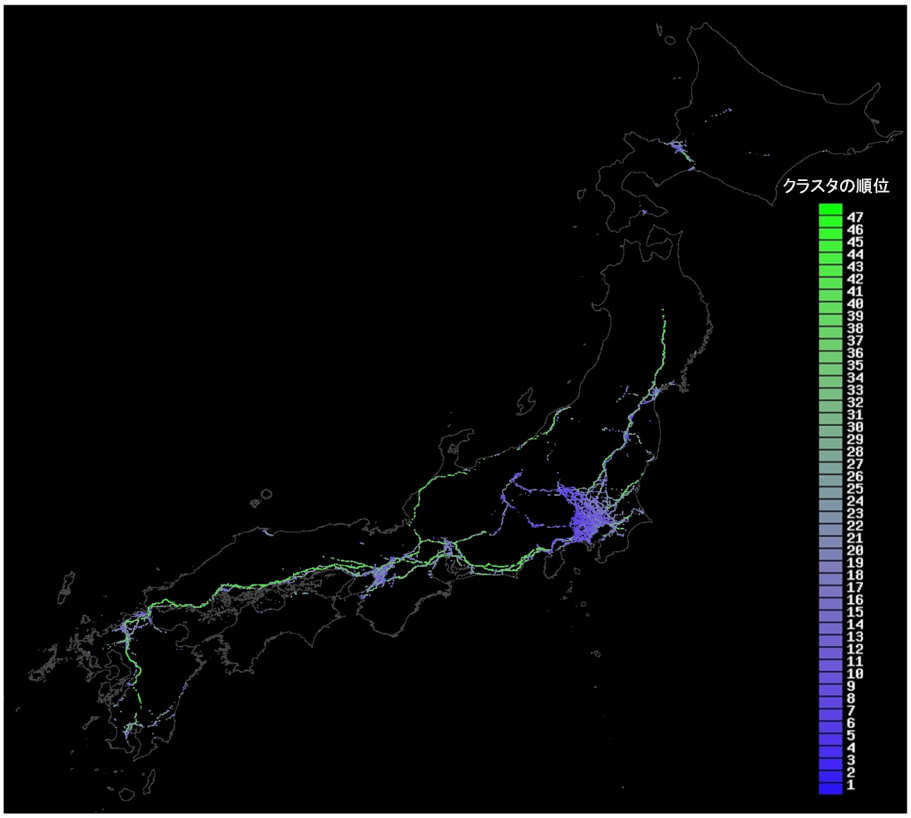

図-5　365日間の速度の日単純平均値の平均順で色づけしたクラスタ分布

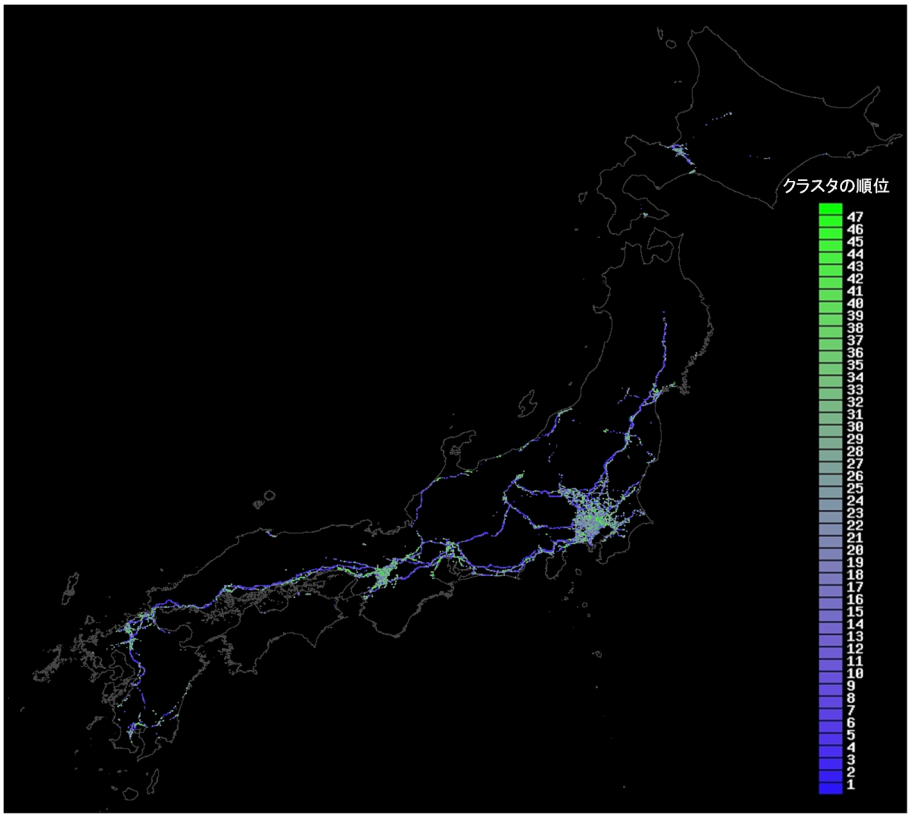

図-6　365 日間の速度の日単純平均値の分散順で色づけしたクラスタ分布

面積（約37万km²）のうち約半数のメッシュに何らかのデータドットデータが含まれていると解釈できる．

図-2は，各メッシュに含まれるドット数の分布である．1ドット以上の観測があったメッシュには，幹線道路だけでなく細街路しかないメッシュやそもそも道路がほとんどないメッシュも含まれているものと考えられ，また，GPSデータのクレンジングを行っていないため，エラーデータのみのメッシュもあると考えられる．このようなことから，ドット数が少ないメッシュから大きいメッシュまで幅広く抽出されている．例えばドット数が多いメッシュに着目すると，観測があったメッシュの12.7%相当する25,009メッシュでは，1年間の間に観測されているドット数が10,000,000ドット以上の観測がされている．これらのメッシュではプローブデータは仕様では1秒毎に観測されていることから，延べ約116日以上に相当の観測があることになる．

以上より，オフピーク等を含む時間帯を継続的に分析することは，観測がない時間帯が多くあり難しいメッシュが多くあることが推察される．一方で，都市部などでは，ある程度の時間・空間分解能を許容すれば，広い範囲で日々の変動などについても継続的な分析ができるものと期待できる．

b) メッシュ化されたデータの可視化例

図-3は，2014/2/13の日平均速度を例として可視化したものである．色で速度を示しており緑が速い速度，青が遅い速度を示している．なお，100,000ドット数が以下だったメッシュについては表示を行っていない．この図によると，比較的交通量が多い都市圏では，色がついているメッシュが多くあること，都市圏外で高速道路を含むメッシュでは速い速度が観測されていることが読み取れ，都市部ではこれらのメッシュに比べ低速となっていることが読み取れる．

図-4は，豪雪があった2014/2/14の日平均速度を災害時の例として可視化したものである．前日と比べ，首都圏全体や宮城県での速度が低下しているほか，山梨県周辺データ観測がないリンクも増えていることが読み取れる．

(3) クラスタリング

a) クラスタリングの実行

2章2節で述べた，クラスタリングの第一段階では，領域を標準地域メッシュの第3次メッシュとし，各領域に対して，集計時間数$T = 6$，すなわち$T \times w = 30$分として，$V_{A,D}$を作成した．クラスタ数は，$k_1 = 2$個から$k_1 = 33$個を試し，Dunn指標[9]が最も小さくなった33個を採用した．なお，クラスタ数を増やした場合はさらに小さいDunn指標が得られる可能性があり，今後検討する．

第二段階では，ラベルデータ①のラベルを，各領域毎に365日分並べたものをL_Aとした．クラスタ数は，$k_2 = 2$個から$k_2 = 48$個を試し，Dunn指標が最も小さくなった48を採用した．

b) クラスタリングの結果例

図-5は，第二段階目のクラスタ結果を集計し速度の日単純平均値を年間で平均したものを算出し，各クラスタの日単純平均速度の大きさの順位によって基づいて着色したものである．緑は平均が大きいクラスタである．前節で示したある平日に観測された傾向と同様に都市圏外で高速道路を含むメッシュでは速い速度が観測されていることが読み取れ，都市部ではこれらのメッシュに比べ低速となっていることがわかる．**図-6**は，速度の日単純平均値の年間での分散をその大きさの順位によって基づいて着色したものである．緑は分散が大きいクラスタである．この図より，都市部のメッシュは，分散が比較的大きなクラスタに分類される傾向が読み取れる．

4. まとめ

本研究では，プローブデータを時空間メッシュごとのデータとして集計した中間データを生成する方法を提案し，さらにデータのクラスタリングの方法を用いたベクトル量子化を用いる中間データの生成する方法を提案した．前者は，生成したメッシュ・時間単位を最小単位とした任意のエリアや時間単位で，プローブカーの流入量，流出量，発生交通量，集中交通量及び，プローブカー平均交通量，平均旅行速度を再集計可能なの中間データとして生成し，集計単位の異なるデータ分析を効率的に行うことを意図するものであった．後者は，前者のデータを前提とした方法であり，集計値の整合性が犠牲にはなるが，蓄積されたデータの集計の試算等を瞬時に行うOLAP (Online Analytical Processing)の要素がある分析を行うことを意図したものである．本研究では，これらの方法を，全国，365日間収集されたプローブデータに適用した．本稿では，単純な適用結果を例示したが，今後の分析では，提案手法の検証を行う予定である．例えば，クラスタリングを用いた方法では，集計値の整合性を犠牲にした方法であることから，精緻に計算した指標値との差異がどの程度あるかということや，計算速度について，定量的に評価する必要がある．

謝辞：本研究で用いたプローブカーのデータは，（株）富士通交通・道路データサービスより提供していただいたものである．ここに感謝の意を表します．

参考文献

1) Zito, R., Este, G. D., Taylor, M. A. P. : Global position-ing systems in the time domain: How useful a tool for intelli-

gent vehicle-highway systems?, Transportation Research Part C, Vol.3, pp.193–209, 1995.
2) Sermons, M. W. and Koppelman, S. : Use of vehicle positioning data for arterial incident detection, Transportation Research Part C, Vol.4 (2), pp.87–96, 1996.
3) 田中康仁，小谷通泰，中村賢一郎：プローブデータを活用した貨物車による配送活動の実態分析，土木計画学研究・論文集，Vol.22，pp.715‐722，2005.
4) 三輪富生，森川高行：プローブカーデータを利用した経路選択行動に関するモデル分析，土木計画学研究・論文集，Vol.21，, pp.553‐560，2004.
5) 宇野伸宏，永廣悠介，飯田恭敬，田村博司，中川真治：バスプローブデータを利用した所要時間信頼性評価手法の構築，土木計画学研究・論文集，Vol.23，pp.1019‐1028，2006.
6) 濱島光宏：商用車プローブデータの収集と活用可能性（特集 ビッグデータ），交通工学，Vol.50(1)，pp.30–33，2015.
7) 太田恒平，大重俊輔，矢部努，今井龍一，井星雄貴：携帯カーナビのプローブ交通情報を活用した道路交通分析，土木計画学研究発表会・講演集，Vol.47，323，2013.
8) Arthur, D., Vassilvitskii, S. : k-means++: the advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035, 2007.
9) Dunn, J. C.: Well-separated clusters and optimal fuzzy partitions, Journal of Cybernetics, Vol. 4(1), pp. 95–104, 1974.

（2016.4.22 受付）

CLUSTERING METHOD FOR PRELIMINARY PROCESSING OF PROBE VEHICLE DATA

Takahiko KUSAKABE and Yasuo ASAKURA