社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－MBL－39（4）
2006－ITS－27（4）
2006／11／16

特徴空間の動的構成によるプローブデータのリアルタイム補完技術

蛭田　智昭　　熊谷　正俊　　谷越　浩一郎　　横田　孝義

株式会社日立製作所　日立研究所　〒319－1292　茨城県日立市大みか町　7－1－1

E-mail:　{tomoaki.hiruta.dp, masatoshi.kumagai.ws, koichiro.tanikoshi.uw, takayoshi.yokota.py}@hitachi.com

要約　現況のプローブデータから、交通情報の相関を表す特徴空間を動的に構成することにより、プローブカーのエリアカバー率に応じて、プローブデータの空間的な欠損を補完する技術について述べる。この特徴空間の動的な構成は、現況データを各基底へ射影、各基底の射影ベクトルのノルムの算出、ノルムの値に応じて基底を選択、選択された基底から特徴空間を構成という 4 つのプロセスから成り立つ。この特徴空間の動的構成による補完技術により、プローブカーが希薄に存在する状況においても、欠損をリアルタイムに補完することが可能になる。

キーワード　プローブカー、欠損値、補完、特徴空間射影

Realtime Imputation Method for Probe Car Data with Dynamic Construction of Feature Space

Tomoaki HIRUTA　Masatoshi KUMAGAI　Koichiro TANIKOSHI and Takayoshi YOKOTA

Hitachi Research Lab., Hitachi Ltd.　7-1-1 Omika, Hitachi-shi, Ibaraki, 319-1292 Japan

E-mail:　{tomoaki.hiruta.dp, masatoshi.kumagai.ws, koichiro.tanikoshi.uw, takayoshi.yokota.py}@hitachi.com

Abstract　This paper discusses realtime imputation method for probe car data with dynamic construction of feature space. This method provides traffic information with no missing data even under sparse probe car condition. Feature space has multiple bases which express correlation of a lot of links. This method consists four major processes: feature space projection of current probe data; calculation of projection norm of each basis; selection of bases according to the projection norm; and dynamic construction of feature space. We evaluate the effectiveness of this method with taxi probe data.

Keyword　Probe Car,　Missing Data,　Imputation,　Feature Space Projection

1. 緒言

近年、プローブ交通情報システムが国内外問わず注目されている。このシステムは、車両自身が交通情報収集のセンサとして振舞い、プローブカーと呼ばれる車両が走行した位置情報、時刻情報などの履歴データを収集するものである。収集された履歴データは交通情報センタにアップリンクされ、交通情報に変換され、提供される。このシステムの利点は、路上センサなどのインフラの必要が無く、低コストで広範囲の交通情報を取得できる点にある。

しかしながら、プローブデータを路上センサと同様に扱う場合、データの補完手段が必要になる。なぜならセンサであるプローブカーの走行経路は確率的なものであり、その情報品質は路上センサで収集される連続的な情報とは異なり、空間的・時間的に大きな欠損を含むためである。例えば、プローブカーの台数を全国で 10 万台とした場合、プローブデータが取得できる時間密度は、道路リンク当たり 1 時間に平均 1 回程度である[1]。このプ

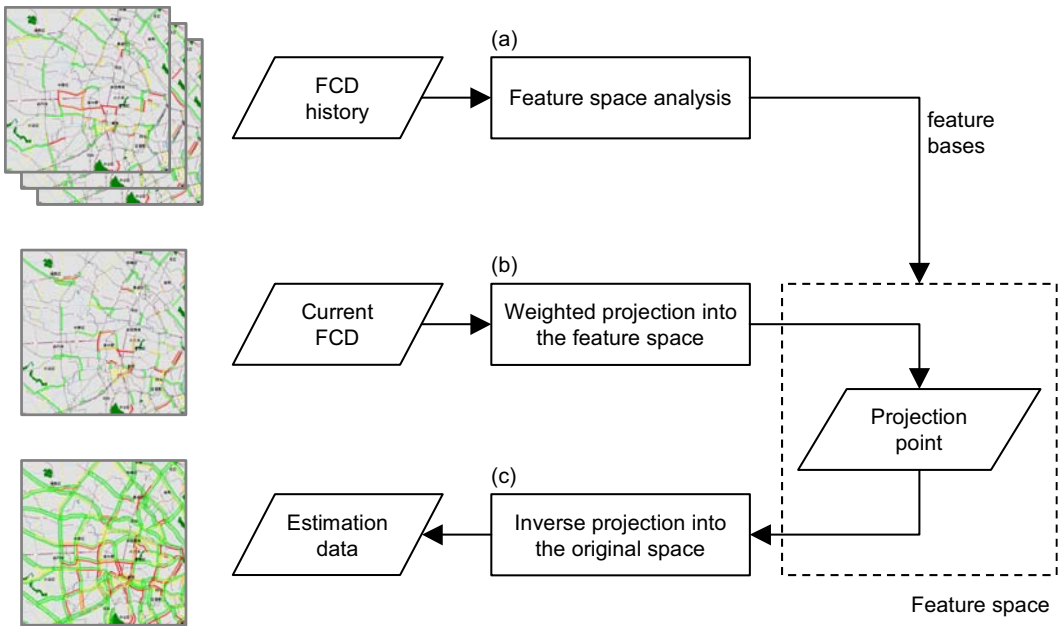

Fig. 1 Process of the realtime imputation

ローブデータを現行の路上センサと同等の 5 分周期のデータとして利用する場合、同時刻でのデータの欠損率は全体の 9 割に達する。よって路上センサと同様に扱う場合、欠損しているデータの補完手段が必要になる。

補完手段の一般的な手法として、過去のプローブデータの同時刻平均値を補完データとして提供する手法がある。しかし、この手法は安定した補完情報を提供することはできるが、曜日や季節の変化に十分に対応できない。また過去データを、曜日、季節のように詳細に分類して、それぞれについて同時刻平均値を求める手法も考えられるが、分類単位ごとにサンプル数が少なくなり、統計的な信頼性は低下する。

この解決策として、特徴空間を用いたプローブカーデータのリアルタイム補完技術が報告されている[2]。この技術は、過去のプローブデータから特徴空間を生成し、現況のプローブデータをその欠損に応じて特徴空間に射影することで、欠損値の補完を行う。この特徴空間は道路リンク間の交通情報の相関関係を表している。

このリアルタイム補完技術では、プローブカーのリンクカバー率（全リンク数に対するプローブ交通情報が収集できたリンク数の割合）が 20%程度あれば、有効な補完を行うことができた。しかし、地方都市などプローブ情報の収集が困難な地域では、リンクカバー率20%の達成も容易ではない。このため、さらに希薄なプローブ情報から、交通情報の空間的な補完を行う必要がある。しかし、例えばプローブカーのリンクカバー率 5%の地域に従来のリアルタイム補完技術を適用すると、補完結果が不安定になるという問題点があった。

そこで本報告では、プローブカーのリンクカバー率が極端に低い場合でも安定にリアルタイム補完を適用することを目的とし、現況のプローブ交通情報に合わせて動的に特徴空間の基底を選択する特徴空間基底選択手法をについて述べる。

以下、2 章ではベースとなるリアルタイム補完技術に関して、基本的なアルゴリズムを説明する。3 章では、プローブカーのエリアカバー率が低い場合への拡張を目指して、動的に特徴空間を構成する手法について述べる。4 章では、その効果を検証する。5 章は結言であり、今後の課題、展望について述べる。

2. 特徴空間射影を用いたリアルタイム補完

2.1. リアルタイム補完の基本アルゴリズム

ある単位エリアで収集されたプローブデータを、リンク単位の旅行時間データなどに加工した上で、主成分分析を行う。これにより、複数のリンクのデータを、相関をもって変化する成分と、無相関に変化する成分に分解できる。

さらに、相関のある成分ごとに、単一の代表変量で表すことが可能になるため、データの次数が縮退される。本来の旅行時間データは、前記代表変量を係数として、リンク間の相関関係を表す基準パターン(これを基底と呼ぶ)を線形合成することにより、近似的に表される。このように集約された情報表現が、特徴空間射影である。基底は特徴空間を構成する静的なパラメータであり、前記代表変量が、特徴空間上で動的に変化する座標に対応する。

逆に、現況の交通情報がプローブデータのように大きな欠損を含むものであっても、それを特徴空間に射影することができれば、その特徴空間座標を元の交通情報データ空間に逆射影することで、交通情報の欠損したリンクについて推定補完を行うことができる。

以上より特徴空間補完は、Fig. 1に示すように、

(a)過去のデータから特徴空間を生成し、

(b)リアルタイムに観測されたデータから、特徴空間上の座標を定め、

(c)特徴空間座標の逆射影によって、推定情報を生成する、

という3つのプロセスから成り立っている。以下、それぞれのステップについて具体的に説明する。

ステップ(a)

特徴空間の生成には、「欠損値付き主成分分析(PCAMD)」[3][4][5]を用いた。これはプローブデータは大規模な欠損を含むため、通常の主成分分析は適用できないためである。

補完対象エリアにおける M 本のリンクについて、N 回にわたって計測された交通情報データを $N\times M$ 行列 $\boldsymbol{X}$ で表すものとする。$\boldsymbol{X}$ の i 行目の成分を対角要素とするデータ行列 $\boldsymbol{D}_{xi}$、重み行列 $\boldsymbol{V}$、$\boldsymbol{V}_0$ に対して、PCAMDはフロベニウスノルム

$$J=\sum_{i=1}^{N}\mathrm{SS}\left(\boldsymbol{Y}_i-\boldsymbol{e}_M\boldsymbol{u}_i{}^{\mathrm{T}}\right)_{D_{wi},I} \tag{1}$$

$$\boldsymbol{Y}_i=\boldsymbol{D}_{xi}\boldsymbol{V}+\boldsymbol{V}_0 \tag{2}$$

を最小化する問題である。この問題を解くことで、処理対象の交通情報データ $\boldsymbol{X}$ の観測値を、誤差ノルム最小で近似できる複数の基底が得られる。すなわち、交通情報データ$\boldsymbol{X}$を、PCAMDで得られた基底で張られる特徴空間に射影すれば、その逆射影によって与えられるデータは、元の交通情報データに対する最尤推定となる。このとき、特徴空間を構成する基底数を次数と呼ぶ。

ステップ(b)

ステップ(a)で得られた基底に対して、欠損のない現況データを射影する場合には、基底と現況データの内積によって、特徴空間座標は一意に決定される。一方、現況データが欠損を伴う場合には、内積による射影は不可能であり、重み付け射影と呼ばれる次式の解法を用いる。

$$\boldsymbol{a}=\left(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{P}\right)^{-1}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{x}^{\mathrm{T}} \tag{3}$$

ここで、$\boldsymbol{P}$ はPCAMDで得られた基底を並べた行列であり、$\boldsymbol{W}$ は重み付けの行列である。欠損を含む現況データ $\boldsymbol{x}$ に対して、射影点 $\boldsymbol{a}$ が得られる。重み付け射影では、観測データの重みを1、欠損データの重みを0として扱うことで、欠損データのリンクを無視し、現況データが観測されたリンクについて、特徴空間上の射影点と、射影前のデータの誤差ノルムが最小化されるように、射影点を決定する。すなわち、重み付け射影によって得られる特徴空間座標は、観測データに対する最尤推定値である。

ステップ（c）

ステップ（b）の重み付け射影によって得られた特徴空間座標 $\boldsymbol{a}$ を、次式により元のデータ空間へ逆射影する。

$$\hat{\boldsymbol{x}} = \boldsymbol{a}\boldsymbol{P}^{\mathrm{T}} \quad (4)$$

逆射影で得られた $\hat{\boldsymbol{x}}$ は、特徴空間上の射影点が $\boldsymbol{x}$ に対する誤差ノルム最小解であるという性質から、$\boldsymbol{x}$ の観測値に対してはその近似値である。また特徴空間がリンク間の相関関係を表すことから、$\boldsymbol{x}$ の欠損値に対する推定値である。$\boldsymbol{x}$ の欠損値を $\hat{\boldsymbol{x}}$ で置き換えることで、$\boldsymbol{x}$ の補完が為される。

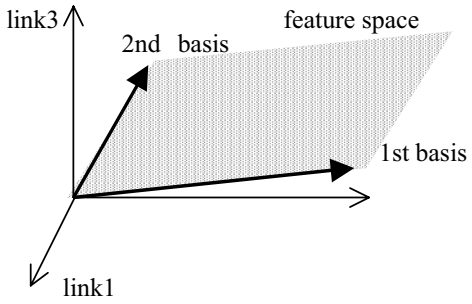

Fig. 2　Example of feature space

2.2. 希薄状況下における問題点

プローブカーが希薄に存在する状況において、前節で解説した特徴空間補完の問題点について述べる。現況のプローブデータを収集できたリンク（以下、プローブデータ観測リンク）数が極端に少ない場合、理論的に推定補完の結果を求めることができない、または推定補完の結果が不安定になるという問題点が生じる。以下、プローブデータ観測リンク数と特徴空間の次数との大小関係を場合分けし、プローブデータ観測リンク数が極端に少ない場合の問題点について説明する。

プローブデータ観測リンク数よりも特徴空間次数が多い場合

理論的に推定補完の結果を求めることができない。これは式(3)の行列 $\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{P}$ が縮退し、その逆行列を導出できないためである。

プローブデータ観測リンク数よりも特徴空間次数が少ない場合

理論的に推定補完の結果を導出できるが、異常な現況データが入力されると、そのデータに全リンクの出力結果が大きく影響し、推定補完の精度が悪くなる可能性がある。高精度な推定補完を実現するためには、特徴空間次数よりも、プローブデータ観測リンク数がある程度以上、多い必要がある。

前述した特徴空間補完では、特徴空間の構成をオフラインで行っていた。このため、特徴空間の次数はオフライン計算時に決定される。具体的には、式(1)から導出した複数の基底の寄与率の高い順に基底を抽出する。この各基底の寄与率は基底の情報量を表しており、上位の基底は補完対象とするリンクにおいて主要な交通情報の変化を表し、下位の基底は軽微な交通情報の変化を表している。

このため、軽微な交通情報の変化までをリアルタイムで補完したい場合は、オフライン計算時に高い次数の特徴空間を構成する必要がある。しかし、現況のプローブデータ観測リンク数が非常に少ない場合には、推定補完結果が不安定または導出できない可能性がある。

この問題を回避するために、オフライン計算時に低い次数の特徴空間を構成した場合、現況のプローブデータ観測リンク数が多い場合であっても、主要な交通情報の変化だけしか補完できず、十分に現況のプローブデータを活かすことができない。さらに、プローブデータ観測リンク数が少ない場合にも、そのリンクに相関のある情報が、少ない次数の特徴空間で表されるとは限らない。

3. 特徴空間の動的基底選択手法

3.1. 基本概念

現況データの相関の強い基底をリアルタイムに選択し、特徴空間を動的に構成することにより、プローブカーのエリアカバー率に応じて、プローブデータの空間的な欠損を補完する手法を述べる。

特徴空間を動的に構成するとは、前節2.1 のプロセス(a)で算出される複数の基底を、現況のプローブデータに合わせて選択し、特徴空間を構成することである。

基底は、プローブデータ観測リンクの相関を強く表しているものを選択する。特徴空間の一例を Fig.2 に示す。基底は各リンクにおいて相関を持って変化する交通情報の成分で構成される。例えば基底１におけるリンク１、リンク２、リンク３のそれぞれの成分を [l_{11}, l_{12}, l_{13}]=[*0.1, 1.0, 0.2*] とすると、リンク１～３の交通情報に”１：１０：２”という比例関係で変化する成分が含まれていることを意味している。基底２の成分は [l_{21}, l_{22}, l_{23}]=[*0.1, 0.2, 1.0*] とする。Fig.2 の基底１は、リンク２の相関を強く表している。このため例えば、基底を１本選択する状況で、リンク２のみ現況プローブデータを収集できた場合は、基底１を選択し、特徴空間を構成する。本報告では、現況のプローブデータの相関の強さを、現況データを各基底への射影した射影ベクトルのノルムで評価する。

3.2. 特徴空間の動的基底選択のプロセス

動的な基底選択手法は、

(i)現況プローブ情報を各基底に射影、

(ii)射影した射影ベクトルのノルムを、各基底の分散で重み付けをし、各基底の評価値を算出、

(iii)その評価値の高い順に、プローブデータ観測リンク数に応じた数の基底を選択、

(iv)選択された基底から特徴空間を構成し、現況データを逆射影して、推定補完情報を生成する、

というプロセスから成り立つ。

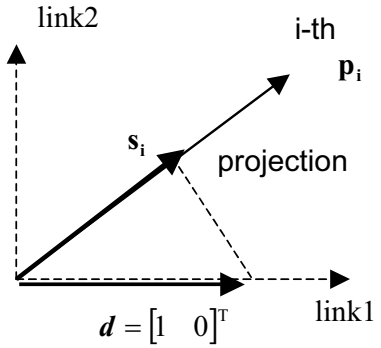

Fig. 3　Evaluation Index.

この基底選択手法は、プローブカーのリンクカバー率に応じて、基底の集合から選択する基底数を可変にする。よって、プローブカーが希薄に存在する状況だけでなく、あらゆるプローブカーのリンクカバー率について適用できる。

カバー率が高い場合は、オフラインで作成した基底の集合から、多数の基底を選択して特徴空間を構成し、推定補完を行う。多くの基底を用いることで、リンク間の軽微な交通情報の変化まで補完でき、精度の高い推定補完を実現できる。一方、リンクカバー率が低い場合は、少数の基底を選択して特徴空間を構成することで、データを収集したリンクの相関の強い特徴空間で推定補完を実現できる。

3.3. 特徴空間の動的基底選択の詳細

現況プローブデータのリンク情報を、各基底への射影する(前節３．２のプロセス(i))。補完対象のエリアにおける M 本のリンクにおいて現況プローブデータのリンク情報ベクトル $\boldsymbol{d}$ を

$$\boldsymbol{d} = [d_1 \quad d_2 \quad \cdots \quad d_M]^T \tag{5}$$

とする。この d_i は、i 番目リンクにおいてプローブ交通情報を収集できた場合は 1、収集できず欠損している場合は 0 の値をとる。例えば、

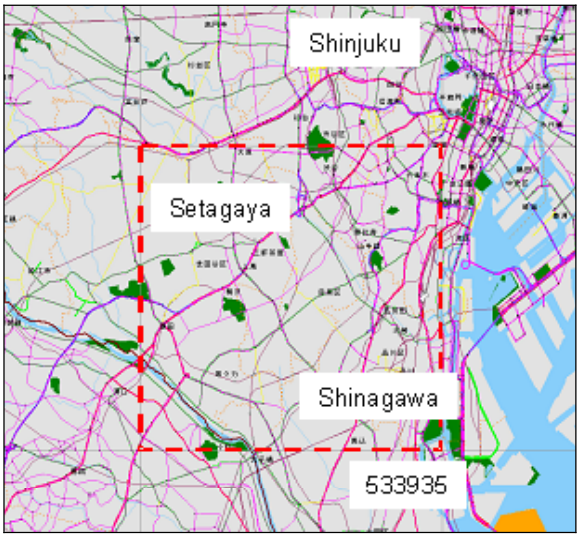

Fig. 4　Evaluation area.

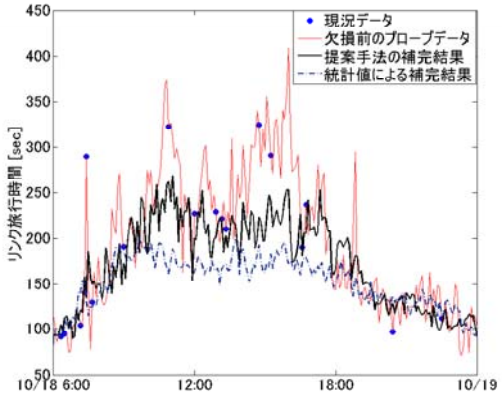

Fig. 5　Example of imputation.

プローブカーからリンク 1、リンク 2 のプローブ交通情報を収集でき、リンク 3 の交通情報が収集できず欠損しているとき、現況プローブデータのリンク情報ベクトルは $\boldsymbol{d}$ =[1 1 0]^{T} となる。

現況プローブデータのリンク情報ベクトル $\boldsymbol{d}$ を i 番目の基底ベクトル $\boldsymbol{p}_i$ へ射影したとき、基底ベクトル空間における射影点座標 t_i は、現況データベクトル $\boldsymbol{d}$ との内積から、

$$t_i = \boldsymbol{p}_i^T \boldsymbol{d} \tag{6}$$

である。これを元のリンク座標系で表すと、

$$\boldsymbol{s}_i = \boldsymbol{p}_i \boldsymbol{p}_i^T \boldsymbol{d} \tag{7}$$

となり、i 番目の基底ベクトル $\boldsymbol{p}_i$ への射影ベクトルになる。Fig. 3は、リンク1とリンク2のリンク座標系上の現況データベクトル $\boldsymbol{d}$ =[1 0]^{T} を i 番目の基底ベクトル $\boldsymbol{p}_i$ へ射影させた例である。

次に、各基底の射影ベクトルを用いて、各基底の評価値を算出する(前節3．2のプロセス(ii))。前述で求まった射影ベクトル $\boldsymbol{s}_i$ のノルムは、i 番目の基底と、現況プローブデータを収集したリンク群との相関の強さを表している。この射影ベクトルのノルムを用いて各基底の評価値を算出する。i 番目の基底ベクトル $\boldsymbol{p}_i$ の評価値 $\boldsymbol{v}_i$ は

$$v_i = \lambda_i \left\| \boldsymbol{s}_i \right\| = \lambda_i \left\| \boldsymbol{p}_i \boldsymbol{p}_i^T \boldsymbol{d} \right\| \tag{8}$$

とする。λ_i は PCAMD の過程で、基底と対で得られる固有値であり、特徴空間の第 i 軸に沿ったデータの分散を表している。λ_i を正規化した値が基底 i の寄与率であることから、式(8)は、射影ベクトルのノルムを寄与率で重み付けした値である。この評価値を用いることで、各基底と、現況プローブデータを収集したリンク群との相関の強さを評価することができる。この評価値の高い順にN_P個の基底を選択し(前節3．2 のプロセス(iii))、新しい特徴空間を構成する(前節3．2 のプロセス(iv))。N_Pはプローブカーのリンクカバー率により決定される。

4. 特徴空間の動的基底選択手法の検証

4.1. 検証手順

本報告では、実際のプローブデータからリンク旅行時間を人為的に欠損させ、そのデータを用いて提案手法の検討を行う。

（1）　検証には東京都内 2 次メッシュ 533935 内におけるタクシーのプローブデータを用いた。Fig. 4に評価エリアを示す。データの蓄積期間は 1 ヶ月分（2005 年 10 月 1 日～31 日）である。タクシーのプローブデータを地図にマッチングし、5 分おきのリンク旅行時間に変換した。ここで

評価対象のリンクは、2次メッシュ 533935 内の主要なリンク 598 本である。

（2） 2005 年 10 月 1 日から 2 週間分のデータを用いて特徴空間の基底を PCAMD を用いて算出する。また、比較のために、統計値を用いた従来手法として、同期間において同時刻平均を計算する。

（3） 手順（2）で得られた基底を用いて、希薄状況での推定補完を実現する。推定補完対象となるプローブデータは、2005 年 10 月 15 日から 31 日のリンク旅行時間データである。ここでプローブカーの希薄な状況を作り出すために、5 分単位のリンクカバー率が 5%になるように、リンク旅行時間データをランダムに欠損させた。

（4） 手順（3）で作成した希薄なリンク旅行時間データを 5 分毎に推定補完する。入力された希薄なリンク旅行時間データから動的に基底を選択し、特徴空間を構成する。このとき選択する基底数は 5 とした（N_P=5）。構成した特徴空間に現況データを射影し、推定補完結果を算出する。

（5） 手順（4）で出力した補完データの誤差評価を行う。このとき比較対象はランダム欠損前の 2005 年 10 月 15 日から 31 日までのリンク旅行時間データである。一方では、統計値による補完の誤差評価も行い、本報告の技術の推定補完精度との比較を行う。

4.2. 検証結果

Fig. 5 は 2 次メッシュ 533935 内の 1 本のリンクの推定補完結果であり、ランダム欠損させた現況データ、元データ、提案した手法の補完結果、従来の統計値による補完結果を時系列順に表示している。縦軸はリンク旅行時間、横軸は 2005 年 10 月 18 日午前 6 時から 19 日午後 0 時まで時系列データを表している。

2005 年 10 月 15 日から 31 日において、リンク毎で補完誤差の平方二乗平均

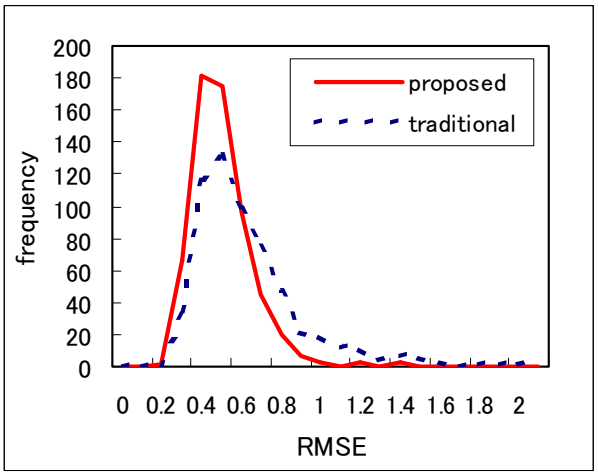

Fig. 6 Evaluation error

（RMSE）を算出した。動的に特徴空間を構成するリアルタイム補完結果の RMSE は 0.45、従来の同時刻平均の統計値による補完結果の RMSE は 0.62 となった。さらに基底選択によるリアルタイム補完結果と統計値の補完結果を比較したリンクの RMSE のヒストグラムを Fig.6 に示す。Fig.6 に示す通り、統計値による補完結果は RMSE が 1.0 を超えた範囲にも広く分布しているのに対し、基底選択によるリアルタイム補完結果は、RMSE が 0.3 から 0.9 の範囲に分布していることがわかる。以上より、絶対的な RMSE の値が大きいものの、リンク毎の RMSE は、従来の補完手法 0.62 から基底選択のリアルタイム補完手法 0.45 へと精度が向上しており、一定の効果が得られたと言える。

RMSE の値が大きい理由は、プローブデータから作成したリンク旅行時間データのばらつきの影響を受けているためと考えられる。Fig. 5 の「欠損前のプローブデータ」で見られるように、真値であるプローブデータはばらついて存在するため、RMSE も大きな値になる。特にリンク旅行時間のばらつきは、元になるプローブカーの台数が少ない場合、少ないプローブカーの挙動に、リンク旅行時間が

影響され、信号の影響等が顕著に現れる。今後プローブデータのばらつきを考慮した誤差評価方法の検討の必要がある。

5. 結言

本研究では、プローブカーが希薄に存在する状況下にて、プローブカーの空間的な欠損を補完することを目的として、特徴空間の動的構成によるリアルタイム補完技術を提案した。また実際のタクシーのプローブデータを用いて、提案した技術の検証を行い、従来の統計値を用いた補完技術よりも補完精度が向上することを確認した。

なお、本報告では過去のプローブ情報を十分に蓄積することで、リンク間の相関関係を表している基底を、十分に算出していることを前提にしている。今後は、プローブカーの過去データから PCAMD を用いて基底を算出する際に、どのくらいの期間の過去データがあれば、プローブカーが希薄に存在する状況でも、十分な基底を算出することができるのか検証する予定である。

一方、現在、日立製作所は、経済産業省の指導で、東京 23 区を対象にしたプローブ交通情報プラットフォームの開発プロジェクト（COSE）にも参加しており、その実用化に当たっても、本研究の成果は寄与できると考える。

6. 謝辞

なお本研究の遂行にあたり、日本交通株式会社殿からタクシーのプローブデータをご提供いただきました。ここに深謝いたします。

参　考　文　献

[1] T. Fushiki, et al., “Study on Density of Probe Cars Sufficient for Both Level of Area Coverage and Traffic Information Update Cycle,” Proc. of 11th World Congress on ITS Nagoya, CD-ROM, Japan, Oct. 2004.

[2] M.Kumagai,et al.,“Spatial Interpolation of Real-Time Floating Car Data Based on Multiple Link Correlation in Feature Space”, Proc. of 13th World Congress on ITS London, CD-ROM, Oct. 2006.

[3] A. Ruhe, “Numerical computation of principal components when several observations are missing”, Tech Rep. UMINF-48,Dept.Information Processing, Umea Univ., 1974.

[4] 柴山、“欠損値がある場合の線形等化法、” 教育心理学研究、 Vol.35、 No.1、 pp.86-89、 1987.

[5] 高根、 “制約付き主成分分析法、” 朝倉書店、 1995.