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

2007－ＩＴＳ－30（４）
2007／9／18

希薄なプローブデータを用いたリアルタイム推定補完技術

蛭田 智昭† 熊谷 正俊† 鈴木 研二‡ 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
‡株式会社日立製作所 CIS 事業部 〒140-0002 東京都品川区東品川 4-12-6(日立ソフトタワーB 16 階)
E-mail: {tomoaki.hiruta.dp, masatoshi.kumagai.ws, kenji.suzuki.qb, takayoshi.yokota.py}@hitachi.com

あらまし 道路間の相関を表す特徴空間を用いて、希薄なプローブデータの空間的な欠損を補完するリアルタイム推定補完技術について述べる。この技術は、蓄積した過去データから特徴空間を生成するオフライン処理と、現況データから特徴空間上の座標を定め、逆射影により推定情報を生成するオンライン処理で構成される。本報告では、オフライン処理に注目し、特徴空間と過去データの欠損率との関係及び、過去データの蓄積期間との関係を明らかにする。そして、プローブカーの希薄な状況下においても、過去データの蓄積期間を長くすることで、特徴空間を十分に生成できるという見通しを得る。

キーワード プローブカー、欠損値、補完、特徴空間射影

Real-time Imputation Method for Sparse Floating Car Data

Tomoaki HIRUTA† Masatoshi KUMAGAI† Kenji SUZUKI‡ and Takayoshi YOKOTA†

† Hitachi Research Lab., Hitachi Ltd. 7-1-1 Omika, Hitachi-shi, Ibaraki, 319-1292 Japan
‡ Car Information Systems Div., Hitachi Ltd. 4-12-6 Higashishinagawa, Shinagawa-ku, Tokyo, 140-0002 Japan
E-mail: {tomoaki.hiruta.dp, masatoshi.kumagai.ws, kenji.suzuki.qb, takayoshi.yokota.py}@hitachi.com

Abstract This paper discusses real-time imputation method for sparse floating car data with feature space which has multiple bases which express correlation of a lot of links. This method consists of off-line and on-line process: determination of feature space from past floating car history (off-line process); feature space projection of current floating data and estimation of missing data performed by inverse projection from feature space (on-line process). In this paper, in the off-line process, we reveal relation between feature space and missing rate of floating car data, and feature space and accumulation time of past floating car data. And we show that feature space can be determined with enough accumulation time even under sparse floating car data.

Keyword Floating 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、時刻情報などのデータを収集するものである。収集された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み得るためである。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

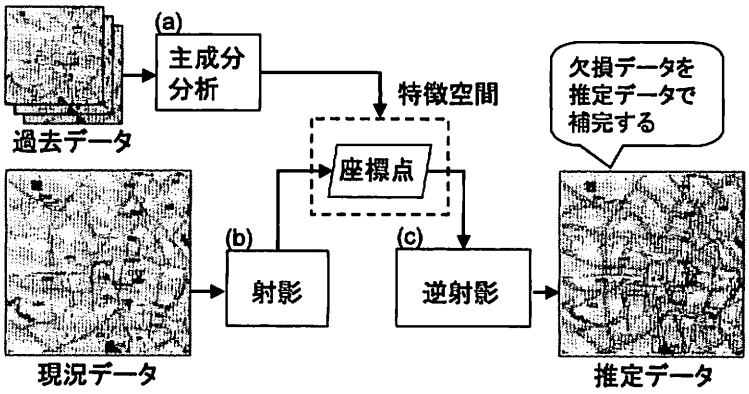

Fig. 1 Process of the realtime imputation

低下する。

この解決策として、特徴空間を用いたプローブデータのリアルタイム推定補完技術が報告されている[2]。これは、道路間の混雑の相関を、特徴空間と呼ばれるモデルで表現し、情報の得られている道路の状況から、他の道路の状況を推定することで、欠損しているデータを補完する技術である。道路間の相関現象は、例えば、幹線道路と駅前の道路が同時に混雑することなどに見られる。この道路リンク間の相関パターンを基底ベクトルと言う。さらに基底ベクトルの集合を特徴空間という。リンク間の相関パターンを抽出するためには、蓄積した過去のプローブデータを用いる。

このリアルタイム推定補完技術では、特徴空間が十分に生成できた場合、プローブデータの欠損率（全リンク数に対するプローブデータが欠損しているリンク数の割合）が 95%以下であれば、安定した推定補完を行うことを筆者らは報告済みである[6]。ここの特徴空間が十分であるとは、抽出した道路リンク間の相関パターンが、実際の道路間の混雑の相関を十分に表すことができる状況をいう。これまでに、蓄積した過去データの欠損率が 80%の場合に、特徴空間を生成した例を報告している[2]。

しかし、地方都市などプローブデータの収集が困難な地域では、欠損率は 80%以上の場合がほとんどとなると報告されている。

そこで本報告では、特徴空間の生成に注目し、その性質を明らかにする。さらに、欠損率が高い状況でも特徴空間を生成する見通しを得る。ここでの特徴空間の性質とは、特徴空間と蓄積した過去データの欠損率との関係及び、特徴空間と蓄積した過去データの蓄積期間との関係である。

以下、2 章ではベースとなるリアルタイム推定補完技術に関して、基本的なアルゴリズムを説明する。3 章では、特徴空間の基底ベクトルと蓄積データの欠損率の関係を明らかにする。4 章では、特徴空間と蓄積データの蓄積期間との関係を明らかにする。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。(a)はオフライン処理、(b)(c)はオンライン処理である。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。

$$\hat{\boldsymbol{x}}=\boldsymbol{a}\boldsymbol{P}^{\mathrm{T}} \tag{4}$$

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。

2.2. 希薄状況下における問題点

リアルタイム推定補完のプロセスは、オフライン処理とオンライン処理からなり、プローブデータの希薄状況下における推定補完では、それぞれ課題がある。

課題１：オンライン処理の課題は、推定結果の不安定性である。これは、希薄なプローブデータを特徴空間に射影するため、欠損データの推定結果が不安定になるという問題であるが、この課題は既に解決済みである[6]。

課題２：オフライン処理の課題は、特徴空間の不完全性である。これは、希薄なプローブデータでは、道路間の相関パターンを十分に生成できず、特徴空間が不完全になるという問題である。

生成した特徴空間が道路間の相関関係を十分に表しているかどうかは、過去データの欠損率とその蓄積期間に依存する。これは、ステップ(a)の欠損値付き主成分分析の手法が、同時刻のプローブデータを収集したリンク同士の相関関係を見て、特徴空間を生成しているためである。例えば、過去データの欠損率が 95% であるエリアの場合、5%という希薄なプローブデータを拠りどころにして、特徴空間を生成しなくてはならない。

本報告では、このオフライン処理の問題に注目し、特徴空間の性質を明らかにして、課題を解決するための見通しを得る。この特徴空間の性質とは、特徴空間と過去データの欠損率の関係及び、特徴空間と過去データの蓄積期間との関係である。

3. 特徴空間と過去データの欠損率の関係

本節では、蓄積した過去データの欠損率と、特徴空間の関係を明らかにする。

蓄積した過去データは式 (2) の行列 $\boldsymbol{X}$ に相当する。この行列は M 本のリンクについて、N 回にわたって計

測されたプローブデータから構成される $N \times M$ 行列である。例えば、情報の更新周期 5 分（1 日 288 個）のデータを 1 ヶ月間蓄積した場合では、X は 8928（288 ×31）×M の行列になる。

3.1. 評価手法

特徴空間と過去データの欠損率との関係を明らかにするために、欠損なしのデータの特徴空間の基底ベクトルを真値として、蓄積期間を固定し、過去データの欠損率を変化させた特徴空間の基底ベクトルを評価する。評価には、相関係数を用い、その値が 1 に近いほど、対象となる特徴空間の基底ベクトルは十分である。相関係数にはピアソンの積率相関係数を用いる。真値の基底の第 j 基底ベクトル $\boldsymbol{p}_j$ と、評価対象の基底の第 j 基底ベクトル $\boldsymbol{q}_j$ との相関係数 r_j を算出する。第 j 基底ベクトル $\boldsymbol{p}_j$ を式 (1) とし、第 j 基底ベクトル $\boldsymbol{q}_j$ を

$$\boldsymbol{q}_j = \begin{bmatrix} q_{1j} & q_{2j} & \cdots & q_{Mj} \end{bmatrix}^T \tag{5}$$

とする。このとき第 j 基底ベクトル $\boldsymbol{p}_j$ と、第 j 基底ベクトル $\boldsymbol{q}_j$ との相関係数は

$$\gamma_j = \left| \frac{\Sigma(p_{ij}-\bar{p}_j)(q_{ij}-\bar{q}_j)}{\sqrt{\Sigma(p_{ij}-\bar{p}_j)^2}\sqrt{\Sigma(q_{ij}-\bar{q}_j)^2}} \right| \tag{6}$$

となる。ここでは、相関係数は絶対値を取るため、γ_j は 0 から 1 の値をとる。

ここでは、特徴空間の基底ベクトルの相関係数が 0.8 以上であれば、道路間の空間的な相関を十分に表現できているとする。

Fig.2 は、特徴空間と、過去データの欠損率の評価手順を示した図である。この詳細を以下に示す。

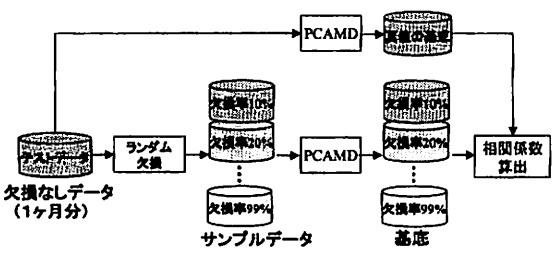

Fig. 2 Evaluation process (1)

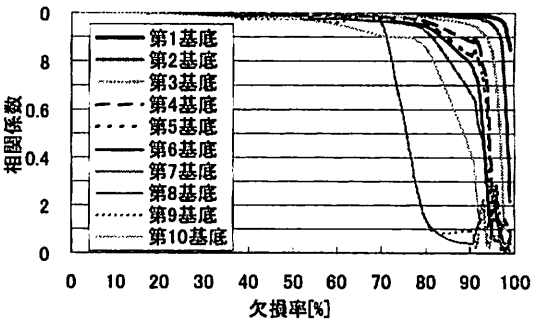

Fig. 3 Evaluation result (1)

- 東京都内品川駅周辺 10km 四方のリンク旅行時間データ（2005 年 10 月分、447 本分）をテストデータとする。このデータは、プローブデータを模擬している。
- テストデータから主成分分析により、真値の特徴空間の基底ベクトルを第 1 から第 447 まで生成する。
- テストデータを欠損率 10%～99%でランダムに欠損させる。
- 欠損データから PCAMD により、特徴空間の基底ベクトルを生成する。
- 真値の基底ベクトルと、欠損データの基底ベクトルの相関係数を算出する。

3.2. 評価結果

Fig.3 は、第 10 までの上位基底ベクトルの相関係数と過去データの欠損率の関係を表している。欠損率 70%までは、相関係数は 0.8 以上であり、十分に基底ベクトルが生成できている。しかし、欠損率 90%を超えると、道路間の相関パターンが十分に抽出できないため、各基底ベクトルの相関係数は、急激に減少する。特に過去データの欠損率が 95%の状況では、第 3 以降の基底ベクトルは不十分になる。推定補完を行う際には、基底の寄与率などの関係から、上位第 5 位の基底ベクトルが最低限必要であるため、欠損率 95%と希薄なプローブデータでは、十分な特徴空間を生成することはできない。

以上の結果より、特徴空間は、過去データの一定の欠損率（ここでは 90%）までは、1 ヶ月の蓄積で特徴空間を十分に生成できることから、欠損率に対して特徴空間の生成はロバストである。しかし、欠損率 90%を超えると、特徴空間は十分に生成することはできない。

4. 特徴空間と過去データの蓄積期間の関係

本節では、欠損率95%の過去データを用い、特徴空間と過去データの蓄積期間の関係を明らかにする。

過去データの蓄積期間は、式（2）の行列X（$N \times M$）の行数Nに相当する。蓄積期間を長くするとは、Nを増やすことである。例えばNの大きさは、蓄積期間1ヶ月では8928　（288×31），3ヶ月間では26784（288×31×3）となる。ちなみに95%の欠損率の交通情報を生成するためには、東京都心部で約100台のプローブカーを必要とする。

4.1. 検証手順

欠損なしのデータの特徴空間の基底ベクトルを真値として、欠損率95%で、蓄積期間を変化させた特徴空間の基底ベクトルを評価する。評価には、式(6)の相関係数を用いる。

Fig.4は、特徴空間と過去データの蓄積期間の評価手順を示した図である。この詳細を以下に示す。

- 欠損率95%のサンプルデータを3ヶ月分作成する。ここでは、1ヶ月分のテストデータを異なるパターンで欠損率95%で欠損させ、3つの1ヶ月分の欠損データを作成する。この3つの欠損データを結合することで、3ヶ月分の欠損率95%のサンプルデータを作成する。
- さまざまな蓄積期間のデータからPCAMDにより、特徴空間の基底ベクトルを生成する。サンプルデータから、蓄積期間1日から3ヶ月のデータを作成し、PCAMDを用いて、特徴空間を生成する。例えば、蓄積期間N日の場合、2005年10月1日から10月N日までのサンプルデータの特徴空間を生成する。
- 真値の基底ベクトルと、欠損データの基底ベクトルの相関係数を算出する。

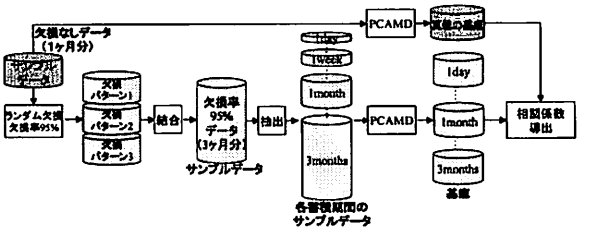

Fig. 4 Evaluation process (2)

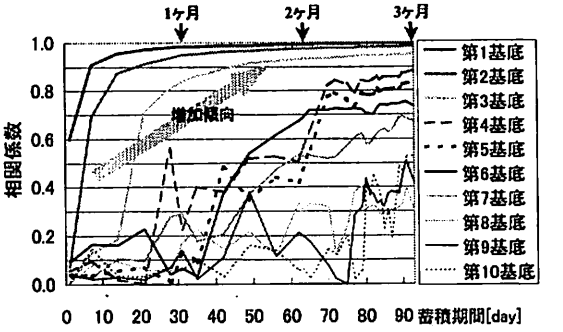

Fig. 5 Evaluation result (2)

4.2. 検証結果

Fig.5は、欠損率95%データの上位基底ベクトルの相関係数と過去の蓄積期間の関係を表している。

前節と同様に、蓄積期間1ヶ月では、十分な基底ベクトルは、第1から第3までである。さらに蓄積期間を長くすることで、各基底ベクトルの相関係数が増加し、蓄積期間3ヶ月では、上位第5基底ベクトルまでは相関係数が0.8以上になり、十分な基底ベクトルを生成できることを確認した。より下位の基底ベクトルも増加傾向を示しており、蓄積期間の延長により十分に生成できると考えられる。

以上の結果から、欠損率95%と希薄なプローブデータでも、過去データの蓄積期間を長くすることで、十分に特徴空間を生成できる見通しを得た。これは、希薄なプローブデータでも、薄く広くとることで、道路間の相関パターンを抽出できるためと考えられる。

5. 結言

本研究では、プローブカーの空間的な欠損を補完するリアルタイム推定補完技術において、オフライン処理で求める特徴空間の性質を明らかにした。この性質とは特徴空間と過去データの欠損率との関係及び、特徴空間と過去データの蓄積期間との関係である。

これにより、欠損率95%の過去データを3ヶ月蓄積することで、第5基底ベクトルまで十分に生成できることから、過去データの蓄積期間を長くすることで、プローブカーが希薄な状況下でも特徴空間を十分に生成できるという見通しを得た。

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.

[6] 蛭田、"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、" 情報処理学会研究報告、第 27 回高度交通システム、 2006.