

多変量拡張軌跡の時空間ペナルティを考慮した類似検索

大橋 英明[†] 清水 敏之^{††} 吉川 正俊^{††}

[†] 京都大学大学院情報学研究科 〒 606 - 8501 京都府京都市左京区吉田本町

^{††} 京都大学大学院情報学研究科 〒 606 - 8501 京都府京都市左京区吉田本町

E-mail: [†]ohashi@db.soc.i.kyoto-u.ac.jp, ^{††}{tshimizu,yoshikawa}@i.kyoto-u.ac.jp

あらまし 近年、移動軌跡に関する研究分野において、単純な軌跡を用いた研究でなく、軌跡と外部情報を組み合わせた研究が注目を集めている。例えば、利用者の行動情報と軌跡を組み合わせることで新たな知識発見を試みた研究や、気象情報、交通量、人の軌跡などを組み合わせることで空気の質の推定を行った研究がある。本研究は、外部情報として数値データの時系列を考え、観測時刻ごとに軌跡上の座標と時系列上の数値データを同時に得られる状況を想定し、軌跡と時系列集合を組み合わせたデータ(多変量拡張軌跡と呼ぶ)を扱う。多変量拡張軌跡の例として、歩行者の軌跡と心拍数の時系列、ハリケーンの軌跡と気圧・風速の時系列などが挙げられる。本研究は動的時間伸縮法(DTW)に基づいた距離尺度を利用した多変量拡張軌跡の類似検索に焦点を当て、ペナルティ関数によって軌跡対のワーピングパスを重視可能な距離尺度を提案する。また、ペナルティ関数を設定する際に利用者の入力が困難と考えられるパラメータの指標も合わせて提案する。

キーワード 軌跡, 多変量時系列, 類似検索

1. はじめに

近年、センシング技術の発達により、膨大な移動軌跡データが収集されている。収集されたデータの例として、歩行者の移動データ[22]、ハリケーンのトラッキングデータ[10]、スポーツのデータ[7]等が挙げられる。また、これらを生の軌跡(raw trajectory)のまま解析するのではなく、データの横断利用が注目される昨今、外部情報と軌跡を組み合わせで解析することで、新たな知見を得ようとする取り組みも盛んである[20]。例えば、位置情報を利用したアプリケーションから得られる利用者の行動情報(外部情報)と利用者の軌跡(軌跡)を組み合わせたデータから頻出パターンを抽出した研究[18]や、気象情報、交通量等(外部情報)と人々の動き(軌跡)を組み合わせ、空気の質を推定した研究[21]がある。前者の研究は外部情報としてラベルの系列データを扱い、新たな知識発見を試みている。後者の研究は外部情報として空気の質に関わる多様なデータを扱い、推定に有効な特徴を設計している。本研究では、外部情報として数値データの時系列を考え、観測時刻ごとに軌跡上の座標と時系列上の数値データを同時に得られる状況を想定する。この想定を満たす例として、歩行者の軌跡と心拍数の時系列、ハリケーンの軌跡と気圧・風速の時系列、サッカー選手の軌跡と選手とボールの位置関係を表す時系列等が挙げられる。本研究は軌跡と時系列集合を組み合わせた上記のようなデータを多変量拡張軌跡と呼ぶ。我々は、動的時間伸縮法(DTW)[3]を用いた距離尺度を利用した多変量拡張軌跡の類似検索に焦点を当てる。多変量拡張軌跡の類似検索はクラスタリングや外れ値検出などの応用手法の基盤になると考えられる。

DTWは時系列の距離算出手法として優れた性能を持つ[15]。また、時系列の伸縮のずれに制限をかける Sakoe-Chiba band [12]を用いた拡張によって、DTWを用いた1-近傍法の精度が上

昇することが報告されている[11]。さらに、Jeongらは Sakoe-Chiba bandを一般化したペナルティを組み込んだ WDTW [9]を提案している。本研究は、Jeongらが提案したペナルティを時間ペナルティと呼び、多変量拡張軌跡の距離尺度に応用することを考えた。

DTWを多変量時系列に適用した簡潔かつ代表的な手法として、Shokoohi-Yektaらが提案した DTW_I , DTW_D がある(I は独立(Independent)を表し、 D は依存(Dependent)を表す)[13]。 DTW_I は各時系列対のDTW距離の総和を多変量時系列間の距離とする手法であり、各時系列対ワーピングパスは異なったものとなり得る。 DTW_D は各時系列対の二点間の距離の総和を用いてDTW距離を求める手法であり、各時系列対のワーピングパスは共通している。また、 DTW_I において、各時系列対のDTW距離の総和を取る際に重み付き線形和を用いる拡張や、 DTW_D において、各時系列対上の二点間の距離の総和を取る際に重み付き線形和を用いる拡張が考えられる。

本研究では、DTWに基づいて多変量拡張軌跡間の距離を算出する際、軌跡対および各時系列対のワーピングパスが大きく異なっていることは不適切であると考えた。例えば、ハリケーンの軌跡と気圧の時系列から構成される多変量拡張軌跡を想定し、多変量拡張軌跡間の距離を算出する場合、気圧はハリケーンの軌跡上の各地点で観測される情報であるため、軌跡対のワーピングパスと気圧の時系列対のワーピングパスが大きく異なっているのは不自然である。この考えに基づくと、単純に DTW_I を用いて多変量拡張軌跡間の距離を算出することは不適切であると考えられる。また、 DTW_D では、ワーピングパスが共通しているが、共通のワーピングパスと軌跡対のワーピングパスが大きく異なる場合がある。そこで、本研究は、軌跡対の二点間の距離に近いほどペナルティが小さくなるようなペナルティ関数を用いて DTW_I を拡張した DTW_P を提案する。

DTW_P は、軌跡対の DTW 距離と、ペナルティが付与された各時系列対の DTW 距離を求め、それらの総和を多変量拡張軌跡間の距離とする。DTW_P を算出するには DTW_D よりも複雑な計算が必要となるが、ペナルティ関数を適切に設定すれば、ペナルティが付与された各時系列対のワーピングパスは軌跡対のワーピングパスと似たものになり、各々のワーピングパスのばらつきが抑えられる。DTW_P では、時間ペナルティの考え方を用いて、ペナルティ関数を実現した。本研究では、軌跡対・各時系列対の DTW 距離の和を取る際に、重み付き線形和を用い、さらに時間ペナルティを付与するような、DTW_P を拡張した手法を DTW_{WTP} と呼び、多変量拡張軌跡の距離尺度として用いることを考えた。DTW_P やその拡張である DTW_{WTP} では、ペナルティ関数を適切に設定するために、いくつかのパラメータを利用者が設定する必要があるが、対象データに対する洞察が不足している利用者にとってはこの設定が困難だと考えられる。そこで、DTW_P のペナルティ関数において特に入力が困難なパラメータの入力支援を目的として、パラメータの指標も合わせて提案した。

本研究の貢献は以下の通りである。

- 多変量拡張軌跡に対して軌跡のワーピングパスを重視可能な距離尺度を提案した
- ペナルティ関数を設定する際に入力が困難なパラメータの指標を提案した

本論文の構成を以下に示す。第 2 節では、関連研究の紹介を行う。第 3 節では、本研究の提案に必要な定義を行う。第 4 節では、DTW_P, DTW_{WTP} を定式化し、第 5 節では、ペナルティ関数を設定する際に入力が困難なパラメータの指標を提案する。第 6 節では、提案した距離尺度の有効性を検討するために行った実験について述べる。第 7 節では、本稿のまとめを述べる。

2. 関連研究

2.1 外部情報と軌跡を組み合わせた研究

ラベルの系列 (利用者の行動情報や地理情報等) と軌跡を組み合わせたデータに対して、効率的な類似検索手法を提案した研究 [19] や頻出パターンを抽出する研究 [18] 等がある。本研究はこれらの研究と異なり、数値データの時系列と軌跡を組み合わせたデータを扱う。上記の研究が扱うラベルは比較の結果として二通りのパターン (一致するか、一致しないか) しか得られないが、本研究が扱う数値データは比較 (差の絶対値を用いた比較など) の結果として多様な値を取り得る。また、ラベルの系列に限らず様々なデータと軌跡を用いて特定のタスクに取り組んだ研究として、気象情報、交通量、人の軌跡などを利用し、空気の質を推定した研究 [21] や、地理情報やタクシーの軌跡から不動産の鑑定を行った研究 [6] がある。本研究はこれらの研究と異なり、(教師付きの) 特定のタスクに取り組むのではなく、多変量拡張軌跡の性質に基づいた汎用的な距離尺度を定式化する。

2.2 時系列の距離尺度に関する研究

本研究は時系列の伸縮を考慮した距離算出手法の一つ、DTW (動的時間伸縮法) を用いる。伸縮を考慮した手法は他にも存在し、LCSS [14], EDR [4] などが挙げられる。これらの中で DTW は最も単純でありながら、優れた性能を持つことが示されている [15]。また近年、ラベル付き時系列データセットの整備が進んだことにより [5], 単変量時系列データ分類する手法が数多く提案されている [1]。時系列の分類問題において、DTW を用いた 1-近傍法は古典的であるが有効な手法とされており、ベースラインとしてしばしば利用されている。本研究は DTW よりも 1-近傍法の精度が良いとされる WDTW [9] を参考にする。WDTW は DTW に対して時系列の伸縮のずれのペナルティを付与した手法である (詳細は後述する)。本研究はこのペナルティを一部拡張したものを時間ペナルティと呼ぶ。本研究は時間ペナルティを応用することで、DTW_P における軌跡対のワーピングパスを重視するためのペナルティ関数を実現する。

2.3 多変量時系列の類似を扱った研究

既存研究として、主成分分析を用いて抽出された特徴量を利用した研究 [16] や、マハラノビス距離と DTW を組み合わせた指標を多変量時系列間の距離とした研究 [17] などがある。多変量時系列は単変量時系列よりも複雑な構造を持つため、他にも多様な類似指標が提案されてきた [2], [8], [13]。本研究は軌跡対のワーピングパスに焦点を当てた距離尺度を提案した点で、多変量時系列の類似を扱った既存研究と異なっている。

3. 各種定義

本節では多変量拡張軌跡を定式化し、DTW および DTW_I に対する時間ペナルティに関して述べる。さらに、DTW_I と DTW_D の定義を述べる。

3.1 多変量拡張軌跡の定式化

我々は軌跡と時系列集合を組み合わせたデータを多変量拡張軌跡と表現する。本研究では軌跡を以下のように表現する。

$$p^i = \langle p^i(1), p^i(2), \dots, p^i(n) \rangle$$

i は軌跡の識別子を、 n は軌跡の観測点の数を、 $p^i(j)$ ($j = 1, 2, \dots, n$) は軌跡上の j 番目の観測点の座標を表している。ただし、本研究ではすべての観測点は等時間間隔で観測されていると仮定する。また、軌跡 p^i と m 個の時系列から構成される時系列集合を組み合わせた多変量拡張軌跡を考える際、時系列集合中の l 番目の時系列を以下のように表現する。

$$v_l^i = \langle v_l^i(1), v_l^i(2), \dots, v_l^i(n) \rangle$$

$v_l^i(j)$ ($j = 1, 2, \dots, n$) は l 番目の時系列中の j 番目の値を表している。以上より本研究は多変量拡張軌跡を以下のように表現する。

$$et^i = \begin{pmatrix} p^i \\ v_1^i \\ \vdots \\ v_m^i \end{pmatrix} = \begin{pmatrix} p^i(1) & p^i(2) & \dots & p^i(n) \\ v_1^i(1) & v_1^i(2) & \dots & v_1^i(n) \\ \vdots & \vdots & \ddots & \vdots \\ v_m^i(1) & v_m^i(2) & \dots & v_m^i(n) \end{pmatrix}$$

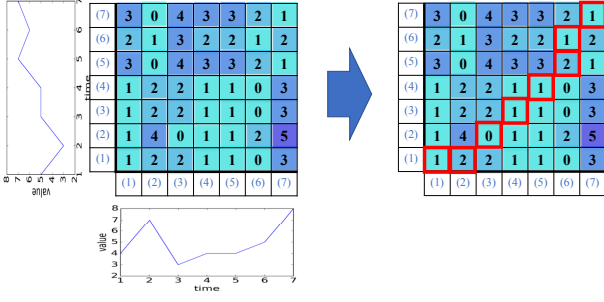


図 1: DTW 概要.

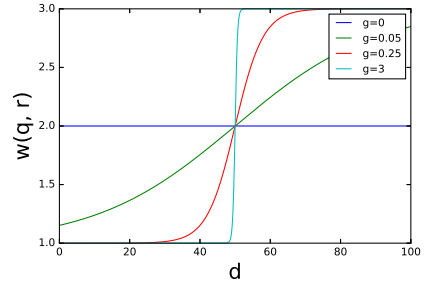


図 2: Modified Logistic Weight Function.

ただし, m は時系列集合の大きさを表している.

3.2 DTW

DTW について述べる. 二つの時系列 $a = \langle a(1), a(2), \dots, a(n_a) \rangle$ と $b = \langle b(1), b(2), \dots, b(n_b) \rangle$ を想定する. このとき, DTW は以下のように定義できる.

$$DTW(a, b) = f(n_a, n_b)$$

$$f(q, r) = D(q, r) + \min \begin{cases} f(q-1, r) \\ f(q, r-1) \\ f(q-1, r-1) \end{cases}$$

$$f(0, 0) = 0, f(q, 0) = f(0, r) = \infty$$

$$(q = 1, 2, \dots, n_a; r = 1, 2, \dots, n_b)$$

ただし, $D(q, r)$ は $a(q)$ と $b(r)$ 間の距離を表している. 図 1 は DTW の概要を表している. 時系列対から導出される距離行列に対して, コストの総和がなるべく小さくなるよう赤色のパス (ワーピングパス) が選択され, パス上のコストの総和が DTW 距離となる. ただし, 図 1 では各時刻における観測値間の距離 ($D(q, r)$) を差の絶対値で算出している. このとき, DTW 距離は $1 + 2 + 0 + 1 + 1 + 2 + 1 + 1 = 9$ となる.

3.3 時間ペナルティ

Jeong ら [9] は DTW に時間の伸縮のずれに関するペナルティを導入した WDTW を提案している. 本研究はこのペナルティを時間ペナルティと呼ぶ. Jeong らは WDTW を定義する際, 先に述べた DTW の定義中の $D(q, r)$ の代わりに, 以下に示す $D'(q, r)$ を用いている.

$$D'(q, r) = w(q, r) \cdot D(q, r) \quad (1)$$

$w(q, r)$ は距離行列中の各要素に付与されるペナルティであり, Jeong らはロジスティック関数を拡張した以下の関数 (Modified Logistic Weight Function) を $w(q, r)$ として用いている.

$$w(q, r) = \frac{w_{max}}{1 + \exp(-g(d - c))} \quad (2)$$

d は比較される点の観測時刻の差 $|q - r|$ である. また, w_{max} は取り得る値の上限を決定するパラメータ, g は関数の形を決定するパラメータ, c は関数の中心を決定するパラメータである. また, すべてのパラメータは 0 以上の値を取る. Jeong らの研究では $g = 0$ (Constant weight), $g = 0.05$ (Linear weight), $g = 0.25$ (Sigmoid weight), $g = 3$ (Two distinct weights) のみが扱われている. $g = 0$ の場合値が一定になり, $g = 0.05$

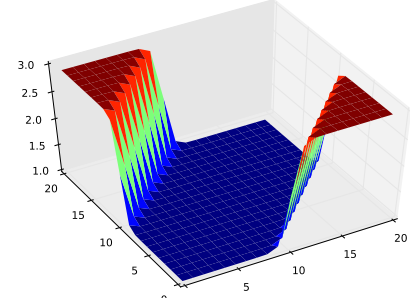


図 3: 時間ペナルティ.

の場合ほぼ線形になり, $g = 0.25$ の場合 c を境に急激に値が変わり, $g = 3$ の場合 c を境に値がほぼ二つに分かれる. $w_{max} = 1, c = 50$ のとき, $g = 0, g = 0.05, g = 0.25, g = 3$ の Modified Logistic Weight Function は図 2 のようになる.

我々は式 2 に柔軟性を与えるため, 下限を調整できるように, 以下のように拡張した.

$$w_t(q, r) = w_{tmin} + \frac{w_{tband}}{1 + \exp(-g_t(d_t - c_t))} \quad (3)$$

d_t は比較される点の観測時刻の差 $|q - r|$ である. w_{tmin} は取り得る値の下限を決定するパラメータ, w_{tband} は取り得る値の幅を決定するパラメータ, g_t は関数の形を決定するパラメータ, c_t は関数の中心を決定するパラメータである. また, すべてのパラメータは 0 以上の値を取る.

$w_{tmin} = 1, w_{tband} = 2, g_t = 3, c_t = 10$ のとき, 長さ 20 の二つの多変量拡張軌跡間の時間ペナルティは図 3 のように表される. 時間ペナルティは単変量時系列に対して提案されたものであるが, 次に述べる DTW_I や DTW_D に適用することも可能である.

3.4 DTW_I と DTW_D

Shokoohi-Yekta らは DTW を用いて多変量時系列データの類似を求める手法を DTW_I と DTW_D の二つに分類している [13]. この二つの手法は DTW を用いた多変量時系列間の距離算出手法として, 簡潔かつ代表的な手法と考えられる. 以下, DTW_I と DTW_D の概要を例を用いて簡単に説明する.

まず, DTW_I について述べる. DTW_I は各時系列対の DTW 距離の総和を多変量時系列間の距離とする手法である. DTW_I を多変量拡張軌跡に適用すると, 軌跡対の DTW 距離と各時系列対の DTW 距離の総和が多変量拡張軌跡間の距離となる (図 4). 各時系列対のワーピングパスは軌跡対のワーピングパスと

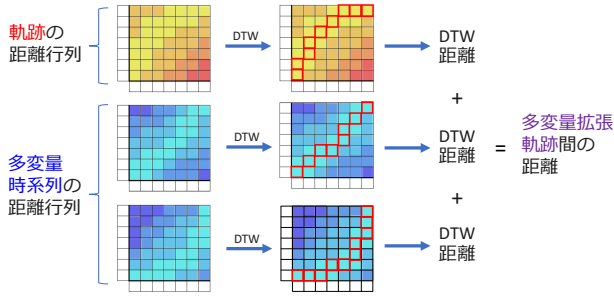


図 4: DTW_I.

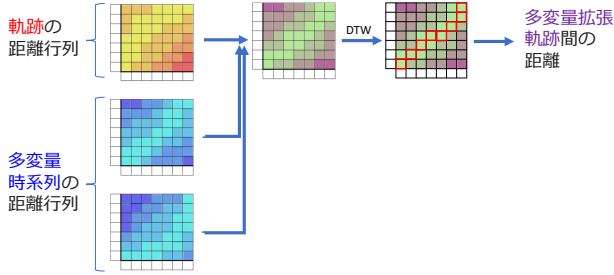


図 5: DTW_D.

異なった経路を取り得るため、DTW_I は多変量拡張軌跡の距離尺度として不適切であると考えられる。

続いて、DTW_D について述べる。DTW_D は各時系列対の距離行列の総和を取って統合した後に DTW 距離を導出する手法である。DTW_D を多変量拡張軌跡に適用すると、軌跡対の距離行列と各時系列対の距離行列を統合した距離行列から導出される DTW 距離が多変量拡張軌跡間の距離となる (図 5)。DTW_D は統合された距離行列に対して単一のワーピングパスを引くため、DTW_I と異なり、軌跡対と各時系列対の間で異なったワーピングパスが取られることはない。しかし、図 5 のように、軌跡対のワーピングパス (図 4 上部) から外れたパスを共通のワーピングパスとして持つ可能性がある。

また、DTW_I において、各時系列対の DTW 距離の総和を取る際に重み付き線形和を用いる拡張や、DTW_D において、各時系列対上の二点間の距離を取る際に重み付き線形和を用いる拡張が考えられる。

4. DTW_P : ペナルティ付き DTW_I

本章は初めに DTW_P の定義を述べる。続いて、DTW_P に重みと時間ペナルティを付与した DTW_{WTP} の概要を述べる。本研究では DTW_{WTP} を多変量拡張軌跡の距離尺度として用いる。さらに、DTW_P のペナルティにおいて入力が困難なパラメータの設定の指標を提案する。

4.1 定義

本研究は軌跡対のワーピングパスを明示的に重視できる DTW_P を新たに提案する。DTW_P は軌跡対の二点間の距離をもとに導出されるペナルティを用いて DTW_I を拡張した距離尺度である。DTW_P は、ペナルティを各時系列対の距離行列に付与し、軌跡対の距離行列とペナルティが付与された時系列対の距離行列をもとに DTW 距離を求め、それらの総和を多変

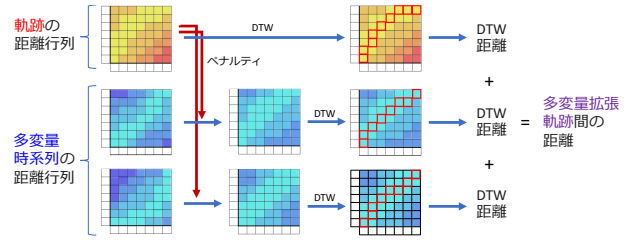


図 6: DTW_P.

量拡張軌跡間の距離とする (図 6)。ペナルティ関数として、軌跡対の二点間の距離が近いほどペナルティが小さくなるような関数を用いるため、DTW_P における各時系列のワーピングパスは、軌跡対のワーピングパスを考慮したものとなり得る。すなわち、各時系列対のワーピングパスが軌跡のワーピングパスと類似したものとなり得る。また、軌跡の影響の大きさは利用者が調整することができ、影響を大きくした場合、DTW_P が描く各時系列対の距離行列のワーピングパスは軌跡対のワーピングパスに近づく。影響を小さくした場合、各時系列対の距離行列のワーピングパスは DTW_I が描くワーピングパスに近づく。影響を完全に無くした場合、DTW_I のワーピングパスと一致する。つまり、DTW_P は DTW_I を一般化した手法となっている。

本研究は、時間ペナルティを参考にし、以下のような関数を DTW_P におけるペナルティとして用いた。

$$w_s(q, r) = w_{s_{min}} + \frac{w_{s_{band}}}{1 + \exp(-g_s(d_s - c_s))} \quad (4)$$

d_s は比較される点間の距離である。本研究では d_s としてユークリッド距離 $\|p^{i_1}(q) - p^{i_2}(r)\|$ を用いる。 $w_{s_{min}}$ は取り得る値の下限を決定するパラメータ、 $w_{s_{band}}$ は取り得る値の幅を決定するパラメータ、 g_s は関数の形を決定するパラメータ、 c_s は関数の中心を決定するパラメータである。また、すべてのパラメータは 0 以上の値を取る。図 7 に示す軌跡対を例に用いて DTW_P におけるペナルティの意義を直感的に述べる。ただし、どちらの軌跡も左下から右上の順番で観測されているものとする。このとき、 $w_{s_{min}} = 1, w_{s_{band}} = 2, g_s = 3, c_s = 60$ ならば、DTW_P におけるペナルティは図 8 のように表され、二つの軌跡上の点が近いほど対応する要素に付与されるペナルティは小さくなり、遠いほど大きくなることを読み取れる。DTW はコストの総和が小さくなるようにパスを選択するため (図 1 参照)、ペナルティを各時系列対の距離行列に掛け合わせることで、図 6 のような軌跡対のワーピングパスを重視したワーピングパスが選択され得る。また、軌跡対のワーピングパスを重視する方法として、単に軌跡対のワーピングパスを用いて他の時系列のワーピングパスを決める方法も考えられるが、この方法の場合、時系列対の距離行列中の大きな値のセルを含むワーピングパスが選択され得る。そのため、本研究は式 4 のようなペナルティを用いた手法がより頑健であると考えた。

DTW_P は、各時系列対の距離行列に対して式 4 のペナルティを与える。すなわち、式 1 中の $w(q, r)$ として $w_s(q, r)$ を用いる。このときの DTW 距離を $WDTW_s(v_l^{i_1}, v_l^{i_2})$ と表現すると、

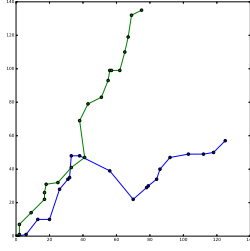


図 7: 軌跡対.

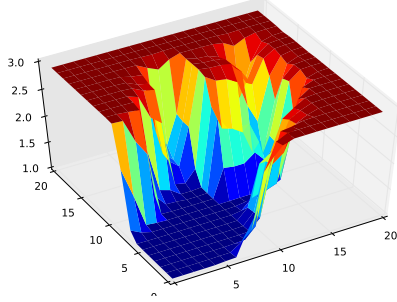


図 8: DTW_P のペナルティ.

DTW_P は以下のように定義される.

$$DTW_P(et^{i1}, et^{i2}) \quad (5)$$

$$= DTW(p^{i1}, p^{i2}) + \sum_{l=1}^m WDTW_s(v_l^{i1}, v_l^{i2})$$

DTW_P におけるペナルティのパラメータについて補足する. パラメータ g_s が 0 のとき (Constant weight), ペナルティの値はすべて等しくなるため, DTW_P のワーピングパスは DTW_I と同じワーピングパスを取る (図 4). すなわち, 軌跡が時系列に与える影響がなくなる. 一方, パラメータ g_s が 3 のとき (Two distinct weights), ペナルティの値は c_s を境に急激に変化するため, c_s が適切に定められかつペナルティ関数の上限値が十分に大きければ, 各時系列対の距離行列のワーピングパスは軌跡のワーピングパスに近づく. すなわち, 軌跡が時系列に与える影響が大きくなる. パラメータ c_s はペナルティ関数の中心を決定する変数であり, どのような値を入力するかで結果は変化する. 5 節では類似検索における c_s の指標を提案する.

4.2 DTW_{WTP}

本研究は, 軌跡対・各時系列対の DTW 距離の和を取る際に, 重み付き線形和を用い, 加えて時間ペナルティを付与し, DTW_P を拡張した手法を DTW_{WTP} と呼び, 多変量時系列間の距離尺度として用いる. DTW_{WTP} では軌跡対の距離行列に対して時間ペナルティ(式 3) のみを付与する. すなわち, 式 1 中の $w(q, r)$ として $w_t(q, r)$ を用いる. このときの DTW 距離を $WDTW_t(p^{i1}, p^{i2})$ と表現する. また, 各時系列対の距離行列に対しては時間ペナルティと式 4 のペナルティを掛け合わせたペナルティを与える. すなわち, 式 1 中の $w(q, r)$ として $w_s(q, r) \cdot w_t(q, r)$ を用いる. このときの DTW 距離を $WDTW_{st}(v_l^{i1}, v_l^{i2})$ と表現する. 以上より, DTW_{WTP} は以下のように定義される.

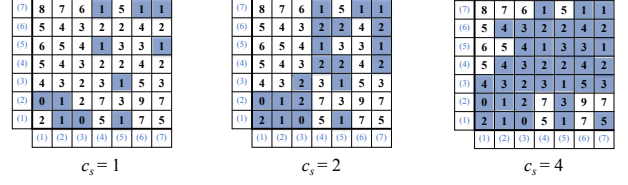


図 9: 軌跡対の距離行列と c_s の関係.

$$DTW_{WTP}(et^{i1}, et^{i2}) \quad (6)$$

$$= W_P \cdot WDTW_t(p^{i1}, p^{i2}) + \sum_{l=1}^m W_{v_l} \cdot WDTW_{st}(v_l^{i1}, v_l^{i2})$$

ただし, $W_P, W_{v_1}, \dots, W_{v_m}$ はそれぞれの距離を足し合わせる際に用いられる重みである. これらの重みは利用者入力を想定している.

5. パラメータに対する指標

本研究はペナルティ関数を持つパラメータの中で特に設定が困難な c_s に焦点を当て, 類似検索において c_s を設定するための指標を提案する. 以下, c_s の設定が困難な理由を説明する. c_s は DTW_P のペナルティ関数の中心を決定する変数であるため, g_s が 0 より大きい場合, 軌跡対の距離行列の中で c_s より大きな要素に対応したペナルティは大きくなり, c_s より小さな要素に対応したペナルティは小さくなる. 図 9 は軌跡対の距離行列に対して, c_s 以下の値のセルを青く塗りつぶした図である. 図 9 に示した軌跡対の距離行列を用いる場合, c_s を大部分の軌跡の観測点間の距離よりも小さくすると, すなわち $c_s = 1$ にすると, 時系列対の距離行列に与えられるペナルティ全体が大きくなってしまふ. 一方で, c_s を大部分の軌跡の観測点間の距離よりも大きくすると, すなわち $c_s = 4$ にすると, 時系列対の距離行列に与えられるペナルティ全体が小さくなってしまふ. よって, 各時系列対の距離行列のワーピングパスが軌跡対の距離行列のワーピングパスに近付くようなペナルティを与えなければ (図 6 参照), c_s を適度な大きさにする必要があると考えられる. しかし, どの c_s が適切なものか, 対象データに対する知見が不足している状況では判断が難しい.

そこで, 本研究は利用者に c_s を直接設定させるのではなく, 直感的に意味を捉えやすい指標 k_c を設定させ, k_c から導出されるパラメータ $c_{min}^{k_c}$ を c_s の代わりに利用するという枠組みを考えた. $c_{min}^{k_c}$ を定義するため, 以下の条件 (A) を導入する.

- (A) ある軌跡対 (p^{i1}, p^{i2}) の距離行列において, c_s 以下のセルのみでワーピングパスを引くことが可能である

本研究は条件 (A) を満たす最小の c_s を $c_{min}^{i1, i2}$ と表現する. まず, 図 9 に示した軌跡対の距離行列を例に用いて説明する. 図 9 において, $c_s = 1$ のとき, 左下端から青く塗りつぶされたセルのみを辿って (右, 上, 右上のいずれかの方向に進むことを繰り返して), 右上端まで到達することはできない. すなわち, 条件 (A) が満たされない. 図 9 の場合, $c_s = 2$ になって初めて条件 (A) が満たされる. よって, $c_{min}^{i1, i2} = 2$ となる. 長さ n_1 の軌跡 $p^{i1} = \langle p^{i1}(1), p^{i1}(2), \dots, p^{i1}(n_1) \rangle$ と長さ n_2 の

軌跡 $p^{i2} = \langle p^{i2}(1), p^{i2}(2), \dots, p^{i2}(n_{i2}) \rangle$ に対して、以下に示す式から $c_{min}^{i1, i2}$ を求めることができる。ただし、ここでは $p^{i1}(q)$ と $p^{i2}(r)$ 間の距離を $D(q, r)$ で表現する。

$$c_{min}^{i1, i2} = h(n_{i1}, n_{i2})$$

$$h(q, r) = \max \left(D(q, r), \min \begin{cases} f(q-1, r) \\ f(q, r-1) \\ f(q-1, r-1) \end{cases} \right)$$

$$(q = 2, \dots, n_{i1}; r = 2, \dots, n_{i2})$$

本研究は、入力軌跡と k_c 件の候補軌跡間の任意の距離行列が (A) を満たすような最小の c_s を $c_{min}^{k_c}$ と表現する。本研究は、直感的に意味を捉えやすい指標 k_c を利用者入力として、 k_c から導出されるパラメータ $c_{min}^{k_c}$ を用いる枠組みにより、適切な値の判断が難しい c_s の設定を支援できると考えた。 $c_{min}^{k_c}$ を求めるアルゴリズムを次に示す。

Algorithm 1 Minimum Warping Threshold Search

INPUT: k_c ▷ the number of candidates which satisfy (A)

INPUT: query trajectory p^q

INPUT: candidate trajectories $p^1, \dots, p^N$

OUTPUT: $c_{min}^{k_c}$

```

1: for  $i \leftarrow 1$  to  $k_c$  do
2:    $PQ.push(c_{min}^{q,i})$ 
3: end for
4: for  $i \leftarrow k_c + 1$  to  $N$  do
5:   if  $c_{min}^{q,i} < PQ.top$  then
6:      $PQ.pop()$ 
7:      $PQ.push(c_{min}^{q,i})$ 
8:   end if
9: end for
10: return  $PQ.top$ 

```

アルゴリズム 1 は k_c とクエリ軌跡 p^q 、候補軌跡 $p^1, \dots, p^N$ を入力とし、 $c_{min}^{k_c}$ を出力とする。まずはじめに、クエリ軌跡と k_c 個の候補軌跡間の $c_{min}^{q,i}$ を計算し、優先度付きキュー PQ に格納する。続いて、それ以外の候補軌跡についても同様に $c_{min}^{q,i}$ を計算し、もし PQ の中で最も大きな値よりも小さければ、その値を PQ から除外し、代わりに $c_{min}^{q,i}$ を格納する。そして最後に、 PQ の中で最も大きな値を出力として返す。

6. 実験

本節では、ハリケーンの実データを用いて、 DTW_P の有効性を検討する。具体的には、 DTW_D にを拡張した DTW_{WD} との比較を行う。

6.1 DTW_{WD} について

DTW_D (図 5) は各距離行列中で同じ位置に存在する要素の単純な和を取り、距離行列を統合していたが、他にも重み付き線形和やマハラノビス距離を用いる方法 [17] などが考えられる。本研究は重み付き線形和を用いて距離行列を統合した場合の DTW 距離を DTW_{WD} と表現する。 DTW_{WD} は DTW の定義において、以下に示す $D_{et}(et^{i1}(q), et^{i2}(r))$ を $D(q, r)$ とし

て用いる。

$$D_{et}(et^{i1}(q), et^{i2}(r))$$

$$= W_p \cdot D_p(p^{i1}(q), p^{i2}(r)) + \sum_{l=1}^m W_{v_l} \cdot D_{v_l}(v_l^{i1}(q), v_l^{i2}(r))$$

ただし、座標間の距離を $D_p(p^{i1}(q), p^{i2}(r))$ 、各時系列上の数値間の距離を $D_{v_l}(v_l^{i1}(q), v_l^{i2}(r))$ 、軌跡と各時系列の重みを $W_p, W_{v_1}, \dots, W_{v_m}$ としている。本研究は DTW_{WD} に時間ペナルティを付与した距離を DTW_{WTD} と表現する。

6.2 実験データについて

本研究はハリケーンのデータを用いて類似検索を行い、空間ペナルティが与える影響に関する考察を行った。ハリケーンの軌跡データとして、Unisys Weather^(注1) の大西洋上のハリケーントラックデータを用いた。本実験では、6 時間ごとに記録された値が欠損していない軌跡データ 451 件を用いる。このトラックデータは軌跡の識別子、タイムスタンプ、緯度、経度、風速、気圧の情報を持つ。ただし、風速と気圧は非常に相関が高いため、本実験では緯度、経度から構成される軌跡に対して風速の時系列データのみを付与した多変量拡張軌跡を扱う。

データの事前処理について述べる。本研究は緯度、経度、風速のデータを各々平均 0、分散 1 となるように正規化した。また、異なる長さの時系列間の DTW 距離よりも、長さが等しくなるように補間された時系列間の DTW 距離のほうが正確であるという知見が既存研究で示されている [11] ため、451 件すべての多変量拡張軌跡の長さが 50 となるよう補間した。

続いて、軌跡間の距離と各時系列間の距離について述べる。すべての距離のスケールを揃えなければ、パラメータの入力意図と出力結果が食い違う可能性があるため、本研究では座標間の距離、時系列間の距離として以下の式を用いた。

$$D_p(p^{i1}(q), p^{i2}(r)) = \sqrt{\frac{(x^{i1}(q) - x^{i2}(r))^2 + (y^{i1}(q) - y^{i2}(r))^2}{2}}$$

$$D_{v_1}(v_1^{i1}(q), v_1^{i2}(r)) = |v_1^{i1}(q) - v_1^{i2}(r)|$$

ただし、 $x^{i1}(q), x^{i2}(r)$ は正規化された緯度、 $y^{i1}(q), y^{i2}(r)$ は正規化された経度、 $v_1^{i1}(q), v_1^{i2}(r)$ は正規化された風速を表している。厳密なスケールの調整は今後の課題とする。

6.3 ペナルティについて

時間ペナルティ・空間ペナルティの設定に必要なパラメータは以下の通りである。

- w_{tmin}, w_{smin} (ペナルティの下限)
- w_{tband}, w_{sband} (ペナルティの幅)
- c_t, c_s (ペナルティの中心を決定する変数)
- g_t, g_s (ペナルティの形を決定する変数)

本実験は $w_{tmin} = 1, w_{tband} = 2, w_{smin} = 1, w_{sband} = 2$ とする。また、 $WDTW_t(p^{i1}, p^{i2})$ 算出の際に、式 1 中の $w(q, r)$ として $w_t(q, r) \cdot w_s(q, r)$ を用い、 $WDTW_{st}(v_1^{i1}, v_1^{i2})$ 算出の際

(注1) : <http://weather.unisys.com/hurricane/>

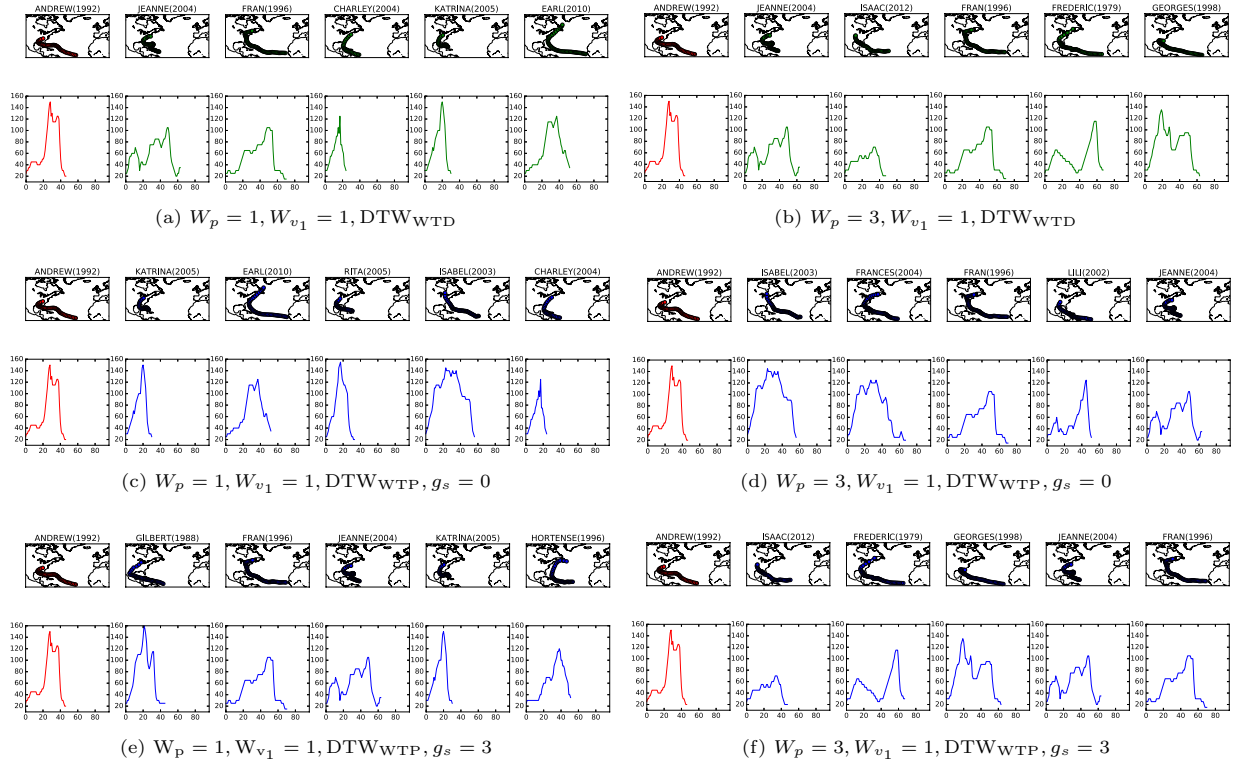


図 10: 実験結果.

に, $w_t(q, r) \cdot w_s(q, r)$ を用いる. これにより, 軌跡間の距離と時系列間の距離に付与されるペナルティの上限はどちらも 9, 下限はどちらも 1 となり, ペナルティによって軌跡間の距離と時系列間の距離のスケールが大きく変わることはないと考えられる.

本実験では $k_c = 50$ としたときの $c_{min}^{k_c}$ を c_s として用いる. また, 二つの多変量拡張軌跡の平均長の 10% の値を c_t として用いる. この値を c_t として選んだ理由を以下に述べる. 時間ペナルティは時系列の伸縮のずれに制限を設ける手法の一つである Sakoe-chiba band [12] の拡張と捉えることができる. よって, Sakoe-chiba band の最適な band 幅がデータセットによって異なるという事実 [11] と同様に, c_t の最適な値はデータセットによって異なると考えられる. しかし実際は, 正解ラベルの存在しないデータセットにおいて c_t の最適な値を求めることは困難である. そこで, 既存研究 [11] 中の「10% を超える band 幅が有効なデータセットの作成ができなかった」という知見を根拠とし, 二つの多変量拡張軌跡の平均長の 10% の値を c_s として用いる. 本実験では $g_t = 3$ とし, 重み, 距離尺度, g_s を変えたときの類似検索の出力結果を考察する.

6.4 実験結果・考察

1992 年にアメリカ合衆国で大きな被害をもたらしたハリケーン ANDREW を入力とし, 重み, 距離尺度, g_s を変化させたときの出力結果を図 10 に示す. 図 10a は $W_p = 1, W_{v1} = 1$ とし, DTW_{WTD} を用いた場合の出力結果である. 図 10b は $W_p = 3, W_{v1} = 1$ とし, DTW_{WTD} を用いた場合の出力結果である. 図 10c は $W_p = 1, W_{v1} = 1, g_s = 0$ とし, DTW_{WTP} を用

いた場合の出力結果である. 図 10d は $W_p = 3, W_{v1} = 1, g_s = 0$ とし, DTW_{WTP} を用いた場合の出力結果である. ただし, 図 10c と図 10d で用いられた距離尺度 ($DTW_{WTP}, g_s = 0$) は DTW_1 に重み (W_p, W_{v1}) を加えた距離尺度と等価である. 図 10e は $W_p = 1, W_{v1} = 1, g_s = 3$ とし, DTW_{WTP} を用いた場合の出力結果である. 図 10f は $W_p = 3, W_{v1} = 1, g_s = 3$ とし, DTW_{WTP} を用いた場合の出力結果である. ただし, 各図の上段がハリケーンの軌跡を, 下段がハリケーンの風速の時系列を, 一番左の列が入力としたハリケーン ANDREW を表している.

図 10a, 図 10c, 図 10e は重みが全て共通 ($W_p = 1, W_{v1} = 1$) しており, 一部の結果は重複しているが, 順序は異なっている. 例えば, ハリケーン KATARINA はすべての出力結果に含まれているが, 図 10c では 1 番目, 図 10a, 図 10e では 4 番目に出現している. KATARINA と ANDREW は出現位置が大幅に異なるが, 部分的に類似した軌跡を辿る, また, ANDREW の風速の時系列は他のハリケーンと比べて, KATARINA の風速の時系列に類似している. よって, 軌跡対と時系列対の DTW 距離はともに近いが, 時系列対ワーピングパスと時系列対のワーピングパスは大きく異なっていると考えられる. 以上より, 軌跡対のワーピングパスと時系列対のワーピングパスの関係を考慮しない $DTW_{WTP}(g_s = 0)$ では 1 番目の出力になったが, 軌跡対のワーピングパスと時系列対のワーピングパスが一致する DTW_{WTD} や時系列対のワーピングパスが軌跡対のワーピングパスに近づくようなペナルティを付与する $DTW_{WTP}(g_s = 3)$ では 4 番目の出力になったと考えられる. 図 10b, 図 10d, 図 10f でも同様に重みは全て共通 ($W_p = 3, W_{v1} = 1$) しており,

一部の結果は重複しているが、順序が異なっている。ただし、図 10a, 図 10c, 図 10e と比べ、時系列の類似が重視されていないことを見て取れる。

ところで、 DTW_{WTD} では W_p を大きくするほど、統合後の距離行列の各要素の値は、軌跡の距離行列の各要素の値に比例した値を取りうる (図 5 参照)。よって、統合後の距離行列のワーピングパスは軌跡の距離行列のワーピングパスに近づくと考えられる。すなわち、多変量軌跡対のワーピングパスが元の軌跡対のワーピングパスに近づくと考えられる。しかし、 DTW_{WTD} は DTW_{WTP} と異なり、軌跡対のワーピングパスのみを重視する手法ではない。ワーピングパスを重視しようとして W_p を大きくすると、軌跡の類似そのものも重視されてしまう。図 10 において、 DTW_{WTD} を用いて軌跡対のワーピングパスを重視しようとしたときの出力 (図 10b) と、 DTW_{WTP} を用いて軌跡対のワーピングパスを重視しようとしたときの出力 (図 10e) を比較する。このとき、 DTW_{WTD} を用いた出力結果の中に、軌跡は類似しているが風速の値が大きく異なったハリケーン (ISAAC) が含まれていることを確認できる。以上より、 DTW_{WTD} を用いて軌跡対のワーピングパスのみを重視することは困難であると考えられるため、軌跡対のワーピングパスのみを重視したい場合は DTW_{WTP} を用いるのが適当であると考えられる。

7. おわりに

本研究は多変量拡張軌跡に対して、軌跡対のワーピングパスを重視するペナルティ関数を用いた距離尺度を提案した。また、ペナルティ関数を設定する際に利用者の設定が困難と考えられるパラメータに対する指標も合わせて提案した。本稿では、実験において、提案した距離尺度をハリケーンの類似検索に適用した際の実例を示すにとどまったが、今後は提案した距離尺度がどのような状況で有効なのか定量的な評価を行いたい。

文 献

- [1] Anthony Bagnall, Aaron Bostrom, James Large, and Jason Lines. The great time series classification bake off: A review and experimental evaluation of recently proposed algorithms. 2016.
- [2] Zoltán Bankó and János Abonyi. Correlation based dynamic time warping of multivariate time series. *Expert Systems with Applications*, Vol. 39, No. 17, pp. 12814–12823, 2012.
- [3] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proc. ACM KDD workshop*, pp. 359–370, 1994.
- [4] Lei Chen, M Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 491–502, 2005.
- [5] Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. The ucr time series classification archive. www.cs.ucr.edu/~eamonn/time_series_data/, July 2015.
- [6] Yanjie Fu, Hui Xiong, Yong Ge, Zijun Yao, Yu Zheng, and Zhi-Hua Zhou. Exploiting geographic dependencies for real estate appraisal: a mutual perspective of ranking and clustering. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1047–1056, 2014.
- [7] Joachim Gudmundsson and Thomas Wolle. Towards automated football analysis: Algorithms and data structures. In *Proc. Australasian Conference on Mathematics and Computers in Sport*, 2010.
- [8] Shen-Shyang Ho, Peng Dai, and Frank Rudzicz. Manifold learning for multivariate variable-length sequences with an application to similarity search. *IEEE transactions on neural networks and learning systems*, Vol. 27, No. 6, pp. 1333–1344, 2016.
- [9] Young-Seon Jeong, Myong K Jeong, and Olufemi A Omitaomu. Weighted dynamic time warping for time series classification. *Pattern Recognition*, Vol. 44, No. 9, pp. 2231–2240, 2011.
- [10] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 593–604, 2007.
- [11] Chotirat Ann Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Proc. ACM KDD workshop*, 2004.
- [12] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, speech, and signal processing*, Vol. 26, No. 1, pp. 43–49, 1978.
- [13] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. Generalizing dtw to the multidimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery*, pp. 1–31, 2016.
- [14] Michail Vlachos, George Kollios, and Dimitrios Gunopulos. Discovering similar multidimensional trajectories. In *Proc. IEEE International Conference on Data Engineering*, pp. 673–684, 2002.
- [15] Xiaoyue Wang, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, Vol. 26, No. 2, pp. 275–309, 2013.
- [16] Kiyoungh Yang and Cyrus Shahabi. A pca-based similarity measure for multivariate time series. In *Proc. the 2nd ACM international workshop on Multimedia databases*, pp. 65–74, 2004.
- [17] Wei Yu and Michael Gertz. Constraint-based learning of distance functions for object trajectories. In *Proc. International Conference on Scientific and Statistical Database Management*, pp. 627–645, 2009.
- [18] Chao Zhang, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta. Splitter: Mining fine-grained sequential patterns in semantic trajectories. *PVLDB*, Vol. 7, No. 9, pp. 769–780, 2014.
- [19] Kai Zheng, Shuo Shang, Nicholas Jing Yuan, and Yi Yang. Towards efficient search for activity trajectories. In *Proc. IEEE International Conference on Data Engineering*, pp. 230–241, 2013.
- [20] Yu Zheng. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology*, Vol. 6, No. 3, p. 29, 2015.
- [21] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-Air: When urban air quality inference meets big data. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1436–1444, 2013.
- [22] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proc. International World Wide Web Conference*, pp. 791–800, 2009.