Data Warehouse de Trajetórias: um Modelo com Suporte à Agregação por Direção dos Movimentos

Carlos Augusto de S. Almeida, Carlos Eduardo Pires, Ulrich Schiel

Departamento de Sistemas e Computação – Universidade Federal de Campina Grande (UFCG)
Caixa Postal 10.106 – 58.429-900 – Campina Grande – PB – Brasil

{carlos, cesp, ulrich}@dsc.ufcg.edu.br

Abstract. *This work proposes a conceptual model for Trajectory Data Warehouses that allows analyzing the behavior of moving objects under and between regions in space and time, according to different levels of granularity, through the use of aggregations. The model enables the segmentation of trajectories into components such as stops and movements. These components can transport semantic information that assign meaning to parts of the trajectory. To reduce the amount of data, the trajectories are stored compactly, summarizing their stops and movements. Experiments were performed to evaluate the level of compaction obtained in the data.*

Resumo. *Este trabalho propõe um modelo conceitual para Data Warehouse de Trajetórias que permite analisar o comportamento dos objetos móveis sobre e entre regiões no espaço e no tempo, de acordo com diferentes níveis de granularidade, através do uso de agregações. O modelo permite a segmentação de trajetórias em componentes, tais como paradas e movimentos. Estes componentes podem transportar informações semânticas que dão significado a partes da trajetória. Para amenizar o problema da grande quantidade de dados, as trajetórias são armazenadas de forma compactada, sumarizando-se suas paradas e movimentos. Experimentos foram realizados para avaliar o nível de compactação obtido para esses dados.*

1. Introdução

A popularização dos dispositivos móveis cientes de localização (*location-aware mobile devices*), tais como telefones celulares e GPS (*Global Positioning System*), possibilitou o monitoramento em larga escala de objetos móveis que transportam esses dispositivos tais como, pessoas, carros e aviões. Esse monitoramento tem como resultado a geração de grandes quantidades de dados sobre as trajetórias desses objetos [Spaccapietra *et al.*, 2008]. A análise desse tipo de dado permite descobrir padrões de comportamento que podem ser explorados em uma grande variedade de domínios [Orlando *et al.*, 2007]. Por exemplo, no gerenciamento de tráfego urbano, a medida "número de veículos que atravessam um cruzamento" é calculada por um sensor que conta os veículos que passam por ele. Essa medida poderia ser substituída por informações detalhadas sobre as trajetórias de cada veículo, incluindo sua origem-destino, rota utilizada, velocidade em cada trecho da trajetória, paradas realizadas, entre outras informações. De forma similar, os dados de trajetórias podem ser usados: no gerenciamento do transporte público, para melhorar a distribuição das linhas de ônibus; no estudo das trajetórias de turistas em uma viagem, para descobrir os locais mais visitados por eles; e no estudo da migração de pássaros, para identificar as rotas migratórias, entre outras aplicações.

A base de dados gerada a partir do monitoramento dos objetos móveis é formada por um conjunto de dados brutos capturados das trajetórias. Para transformar essa massa bruta de dados em informações úteis, uma forma adequada é disponibilizá-la em um *Data Warehouse* (DW), um banco de dados otimizado para lidar com grandes volumes de dados de forma eficiente. Para dados convencionais, DWs têm sido usados com sucesso no decorrer das últimas décadas. Entretanto, a natureza dos dados de trajetória e a grande quantidade desses dados impõem desafios para construção e manutenção do DW, dentre eles: (i) o monitoramento dos objetos móveis gera apenas dados brutos que, para muitas aplicações, não são suficientes para extrair informações úteis. Portanto, antes desses dados estarem prontos para uso, eles precisam ser enriquecidos com informações semânticas [Bogorny *et al.*, 2009]; (ii) o suporte oferecido pelas tecnologias de DW para dados de trajetória ainda está limitado ao armazenamento e recuperação de observações individuais da trajetória [Spaccapietra *et al.*, 2008]. Não existe suporte nativo a trajetórias, como acontece com os dados espaciais; e (iii) a grande quantidade dos dados de trajetória consome muitos recursos, tornando o tempo de processamento das consultas longo, impossibilitando análise no estilo OLAP [Orlando *et al.*, 2007].

É proposto neste artigo um modelo semântico para *Data Warehouses* de Trajetórias (DWTrs) com suporte à agregação por direção dos movimentos. Uma preocupação específica neste trabalho é permitir analisar a movimentação dos objetos móveis "sobre" e "entre" as regiões no espaço e no tempo, **análise orientada a tráfego** e **análise orientada a trajetórias**, respectivamente. A primeira é obtida agregando-se as medidas das trajetórias por espaço e tempo. A análise orientada a trajetórias é proporcionada por um conjunto de dimensões "direção do movimento", que permitem representar as trajetórias sobre diferentes níveis de granularidade. Outra preocupação neste trabalho, é a **modelagem de trajetórias semânticas** [Spaccapietra *et al.*, 2008] em DWTrs. Tal modelagem permite segmentar trajetórias em diversos componentes, tais como *paradas* e *movimentos*, que podem transportar informações que dão significado ao componente que pertencem. Para amenizar o problema da grande quantidade dos dados de trajetória, propõe-se compactar trajetórias mediante a sumarização de suas paradas e movimentos. Dessa forma, consegue-se reduzir drasticamente o tamanho dos fatos, como comprovado através de diversos experimentos realizados.

As demais seções deste artigo estão organizadas como segue. Na *Seção 2* são apresentados os trabalhos relacionados. Na *Seção 3* é descrito o cenário de aplicação usado durante os exemplos deste trabalho. Na *Seção 4* é apresentada a forma adotada para representação de trajetórias. Na *Seção 5* são descritos o modelo proposto, os procedimentos necessários para realizar a carga de dados usando o modelo, e como proporcionar agregação por direção dos movimentos da trajetória. Na *Seção 6* são discutidos os experimentos realizados para comprovar a redução no tamanho dos fatos, parada e movimento. Finalmente, na *Seção 7*, são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Pesquisas sobre trajetórias de objetos móveis são relativamente recentes. Dentre as questões em aberto que vêm despertando grande interesse das comunidades de pesquisa, pode-se mencionar: a *modelagem multidimensional para dados de trajetória* e a *definição e implementação de operadores TrOLAP* (*Trajectory OLAP*, em português *OLAP para Trajetórias*). Os trabalhos de Orlando *et al.*, (2007), Marketos *et al.* (2008) e Baltzer *et al.* (2008) foram pioneiros nesse sentido. Orlando *et al.*, (2007) investigam como armazenar e agregar dados de trajetória usando as tecnologias de DW tradicionais. Para acomodar as trajetórias, o espaço geográfico é dividido em um conjunto de células espaço-temporais. Cada célula armazena, de forma sumarizada, os dados de todas as trajetórias que por ela

passam. Marketos *et al.* (2008) descrevem os procedimentos ETL (*Extraction*, *Transformation, and Load*) [Kimball *et al.*, 2002] necessários para povoar um DWTr baseado no modelo de Orlando *et al.*, (2007). A partir de um conjunto de dados brutos sobre a localização espaço-temporal dos objetos móveis, os autores investigam como extrair as trajetórias dessa base (processo de reconstrução de trajetórias), transformar e carregar os dados. Baltzer *et al.* (2008) propõem um novo operador OLAP para agregação de trajetórias similares, o qual permite identificar objetos móveis que se deslocaram em paralelo.

Os trabalhos envolvendo trajetórias podem ser organizados segundo a classificação oferecida por Andrienko e Andrienko (2008). Segundo os autores, trajetórias podem ser analisadas sob dois pontos de vista: *visão orientada a tráfego* e *visão orientada a trajetórias*, ou, como são chamadas neste trabalho, *análise orientada a tráfego* e *análise orientada a trajetórias*. Na análise orientada a tráfego o objetivo é analisar situações de tráfego, ou seja, analisar o comportamento dos objetos móveis em uma dada região em diferentes intervalos de tempo. Ela é adotada por Marketos *et al.* (2008) e Orlando *et al.* (2007) em seus trabalhos. Na análise orientada a trajetórias o objetivo é analisar o deslocamento dos objetos móveis entre as regiões em termos de origem-destino do movimento. Os trabalhos de Baltzer *et al.* (2008), Gomez *et al.* (2008), Kuijpers e Vaisman (2007), e Spaccapietra *et al.* (2008) consideram esse tipo de análise. Dependendo do modelo adotado, os dois tipos de análise estão disponíveis. Por simplificação, usamos o termo *análise de trajetórias*, para expressar ambos os tipos de análise.

Data Warehouses Espaciais [Bédard *et al.*, 2001] são empregados por Kuijpers e Vaisman (2007), e Gomez *et al.* (2008) para analisar trajetórias. Segundo seus autores, o uso de medidas e dimensões espaciais aumenta o poder de expressividade do modelo, além de simplificar a construção e processamento de algumas consultas. O trabalho de Gomez *et al.* (2008) utiliza a mesma arquitetura usada por Kuijpers e Vaisman (2007), mas seu modelo distingue *paradas* de *movimentos* em trajetórias. Além disso, propõem compactar trajetórias, armazenando apenas as paradas realizadas em cada uma delas, e seus movimentos na forma de transição entre paradas (p.ex. do Banco B1 para o Teatro T1). Conseguem assim, reduzir drasticamente a quantidade de dados armazenada.

Em geral, os trabalhos na literatura permitem apenas *análise orientada a tráfego*, e alguns destes conseguem resolver bem o problema da grande quantidade dos dados de trajetória, como Marketos *et al.* (2008) e Orlando *et al.* (2007). Entretanto, poucos trabalhos proporcionam *análise orientada a trajetórias*, sendo que estes não permitem analisar a direção dos movimentos no estilo OLAP, tais como Baltzer *et al.* (2008), Gomez *et al.* (2008), Kuijpers e Vaisman (2007), e Spaccapietra *et al.* (2008). Dos trabalhos analisados, apenas Gomez, *et al.* (2007) e Spaccapietra *et al.* (2008) distingue paradas de movimentos em trajetórias, o que é fundamental para a análise correta de trajetórias. A inclusão de dados sobre paradas na análise de movimentos pode provocar forte discrepância entre os dados analisados e os reais, e vice-versa. Por exemplo, ao se incluir dados sobre paradas no cálculo da velocidade média de uma região, tem-se uma forte impressão de que a velocidade na região analisada está baixa, devido a influencia dos dados sobre paradas (cuja velocidade é igual a zero) durante as computações.

3. Cenário de Aplicação

Esta seção apresenta um cenário de aplicação usando DW de Trajetórias denominado **gerenciamento de tráfego urbano**. Para essa aplicação exemplo, suponha que uma determinada organização governamental esteja disposta a melhorar o tráfego das cidades que administra. Para isso, essa organização precisa monitorar os indivíduos de uma parcela representativa da população de cada cidade analisada, os quais recebem benefícios do governo para participarem do projeto. Cada indivíduo é monitorado através de seu telefone

celular equipado com um GPS, o qual captura sua localização espaço-temporal a cada 20 segundos. Esses dados são armazenados e posteriormente transmitidos a um servidor.

Para ajudar no enriquecimento semântico das trajetórias, cada indivíduo da população oferece informações detalhadas sobre seu comportamento tais como: (i) informações pessoais: sexo, idade, estado civil, profissão, endereço residencial e comercial; (ii) locais mais frequentados e quando isso ocorre: local de trabalho, casa, bares, escola das crianças; (iii) rotas comumente usadas para ir de um lugar ao outro; e (iv) meio de transporte utilizado. Além disso, são mantidas, informações sobre as cidades analisadas e dados espaciais como: ruas (representadas através de polilinhas), bairros (polígonos) e regiões de interesse (ou RoIs, polígonos representando lugares como hotéis, restaurantes e escolas). Associados a essas regiões podem existir eventos como shows, congestionamentos, alagamentos, acidentes, entre outros.

Para atender parte dos requisitos necessários para o gerenciamento de tráfego urbano, o modelo proposto deverá oferecer as seguintes informações sobre o tráfego de pessoas circulando em uma cidade: **(r_1)** o comportamento dos indivíduos nas regiões, em termos do número de indivíduos, velocidade, locais de parada, entre outras medidas; **(r_2)** a impedância de uma região, ou seja, a obstrução do movimento; **(r_3)** o comportamento dos indivíduos entre as regiões, similar a r_1; **(r_4)** as rotas mais usadas pela população para ir de um lugar ao outro; **(r_5)** os pólos gerados de tráfego; e **(r_6)** a proporção de veículos que deixam uma avenida em suas diferentes saídas [Andrienko *et al.*, 2007; DENATRAN/FGV, 2001]. Os requisitos r_1 a r_2 e r_3 a r_6 são atendidos pela *análise orientada a tráfego* e *análise orientada a trajetórias*, respectivamente.

4. Representação de Trajetórias

Para representar trajetórias, buscou-se um modelo que permitisse distinguir paradas de movimentos em trajetórias, assim como outros componentes. Para isso, o modelo de trajetórias de Spaccapietra *et al.* (2008) foi estendido com algumas modificações em sua face semântica, como discutido a seguir. De acordo com esse modelo, uma trajetória é formada por duas faces: *face geométrica* e *face semântica*.

Por questões de simplificação, a **face geométrica** da trajetória é representada por uma sequencia finita de observações na forma $[((x_1, y_1), t_1), ((x_2, y_2), t_2), ..., ((x_n, y_n), t_n)]$, onde, para cada observação $((x_i, y_i), t_i)$, o par (x_i, y_i) representa a localização espacial, e t_i o tempo, com x_i, y_i e $t_i \in R$, e $t_i < t_{i+1}$. Para reconstruir os movimentos da trajetória entre duas observações consecutivas, é usada a função de *interpolação linear local* (Pelekis *et al.*, 2008), a qual considera que um objeto móvel se desloca em linha reta a uma velocidade constante entre duas observações.

A **face semântica** divide a *face geométrica* em vários componentes, os quais podem transportar informações semânticas definidas pelo usuário, que dão significado à parte da trajetória a qual pertence. Para este trabalho, adotaram-se os seguintes componentes semânticos: *início da trajetória*, *fim da trajetória*, *parada* e *movimento*. Diferente de Spaccapietra *et al.* (2008), considera-se que uma trajetória é delimitada pelo intervalo de tempo $[t_{fim\text{-}traj\text{-}f1}, t_{fim\text{-}traj\text{-}f2}]$, onde $t_{fim\text{-}traj\text{-}f1}$ e $t_{fim\text{-}traj\text{-}f2}$ representam o instante de tempo do fim da parada *f1* e *f2*, respectivamente, como mostrado na Figura 1. Sendo que as paradas *f1* e *f2* ocorrem em regiões de interesse (RoIs).

Por simplicidade, considera-se que as trajetórias capturadas são precisas e que a localização espaço-temporal dos objetos móveis é feita por GPSs, em intervalos de tempo pequenos e regulares (por exemplo, a cada 20 segundos). Os objetos móveis são representados na forma de um simples ponto, que se desloca no espaço e no tempo.

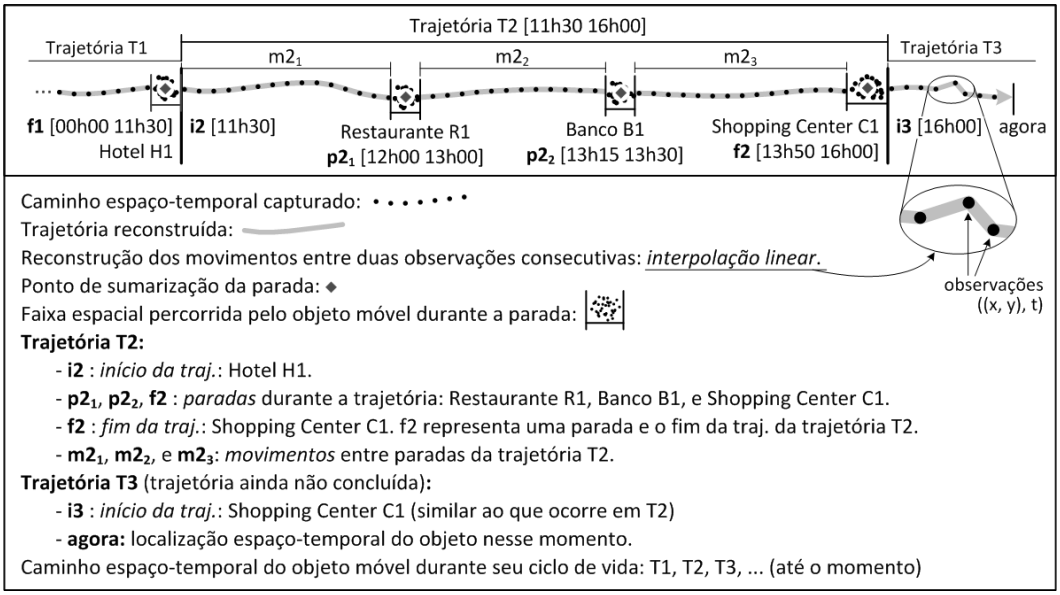


Figura 1. Caminho espaço-temporal de um objeto móvel.

5. Modelo Proposto

O modelo proposto é uma extensão de um DW Espacial [Bédard *et al.*, 2001]. Para representação de trajetórias, o *espaço geográfico* é discretizado por uma grade regular formada por um conjunto de células espaciais, e o *tempo* é discretizado em intervalos de tempo regulares. Em geral, intervalos com duração de algumas dezenas de minutos. Para atingir os objetivos propostos, o modelo adotado incorpora as seguintes dimensões, como ilustrado na Figura 2.

Objeto Móvel: (ObjMovDim) dimensões demográfica e tecnográfica. A *face demográfica* mantém dados sobre os objetos móveis. Por exemplo, no caso de indivíduos, nome, sexo, idade, profissão e estado civil. A *face tecnográfica* mantém dados sobre o dispositivo de localização usado, tais como a precisão do GPS usado.

Trajetória: (TrajDim) dimensão descrita, contém as informações sobre a trajetória como um todo. Basicamente possui as informações: (i) *espaciais*: origem e destino da trajetória; (ii) *temporais*: início e fim da trajetória; e (iii) *descrita*: objetivo da trajetória (por exemplo, indo do trabalho para casa).

Célula: (CelulaDim) dimensão espacial, armazena as células espaciais da grade regular. Em geral, possui a hierarquia *célula* < *bairro* < *cidade*.

Tempo: (TempoDim) dimensão temporal, definida em intervalos de tempo. Mantém os eventos que ocorreram para cada intervalo de tempo como, por exemplo, shows, partidas de futebol, acidentes de trânsito, entre outros.

Região de Interesse: (RoIDim) dimensão espacial, armazena os dados sobre as regiões de interesse (RoIs) tais como nome, categoria (hotel, shopping, universidade, entre outros) e dados espaciais (polígono que representa o RoI). Em geral, possui a hierarquia *roi* < *célula* < *bairro* < *cidade*.

Direção do Movimento: representada pelas dimensões: (i) *DirMovDim*: (direção do movimento entre regiões), mantém a direção do movimento entre as regiões no tempo, através do par origem-destino do movimento (por exemplo, do *bairro1* para o *bairro3*); e (ii) *DirMovAdjDim*: (direção do movimento entre regiões adjacentes) similar a *DirMovDim.*

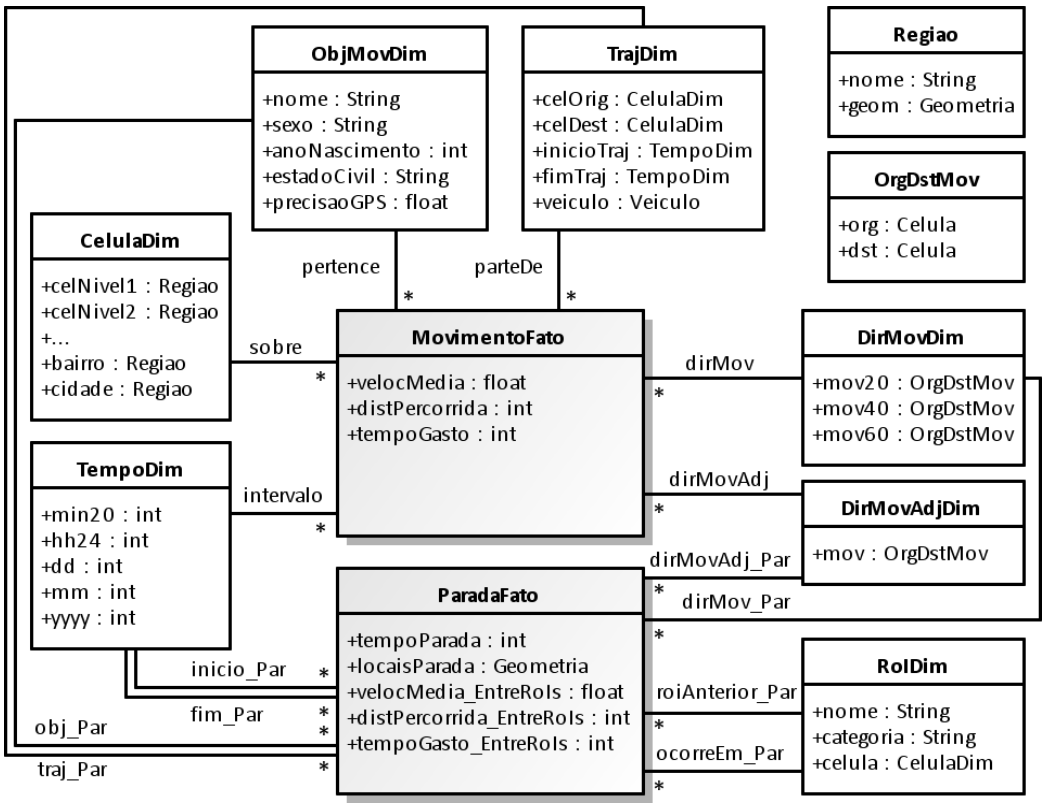


Figura 2. Modelo Proposto ilustrado por um diagrama UML.

A *análise orientada a tráfego* é proporcionada pelas dimensões *CelulaDim* e *TempoDim*, e a *análise orientada a trajetórias* é proporcionada pelo conjunto de dimensões direção do movimento: *DirMovDim*, *DirMovAdjDim* e *RoIDim*. Na Seção 6.3 é descrito o funcionamento da agregação por direção dos movimentos.

Para analisar trajetórias de forma correta, é necessário oferecer uma distinção clara entre paradas e movimentos. Sendo assim, no modelo proposto, são adotados dois fatos: (i) **fato parada** (*ParadaFato*) armazena os dados referentes às paradas na forma sumarizada por parada. Em geral, é representado pelas medidas *tempo de parada* e *local de parada* (um dado espacial, representado por um simples ponto); e (ii) **fato movimento** (*MovimentoFato*) armazena os dados referentes aos movimentos na forma sumarizada por célula espaço-temporal. Representado pelas medidas *velocidade média*, *espaço percorrido*, *tempo decorrido*, entre outras.

5.1. Carga de Dados

Nesta seção são descritos os passos necessários para transformar a sequencia de observações capturadas da trajetória de forma a se adequarem ao modelo proposto. Este trabalho pressupõe que os componentes da trajetória, tais como paradas e movimentos, já foram previamente identificados e as anotações semânticas incluídas. Os passos são os seguintes:

Passo 1 – Sumarização das observações referentes às paradas: as observações referentes a cada parada da trajetória são sumarizadas e armazenadas na forma de um único registro no fato parada. Sendo assim, na Figura 3(a) as observações no intervalo (00h53, 01h33) referentes à parada p1 são sumarizadas e armazenadas como um único registro, como é mostrado Figura 3(b). Para facilitar a compre-

ensão dos exemplos, o identificador das observações da trajetória coincide com o momento de captura da observação.

Passo 2 – Identificar e descartar movimentos dentro de RoIs: a análise dos movimentos da trajetória está interessada nos dados sobre a movimentação dos objetos na célula (ou seja, nas ruas), mas não dentro de regiões de interesse (RoIs). Para distinguir os movimentos que ocorreram *dentro* e *fora* dessas regiões, novas observações são acrescentadas à trajetória, nos pontos de intersecção dela com as bordas espaciais dos RoIs. Por exemplo, na Figura 3(a) é acrescentada a observação 00h14 ao intervalo (00h13, 00h33) para dividi-lo em (00h13, 00h14) e (00h14, 00h33), movimentos *dentro* e *fora* do RoI H1, respectivamente.

Passos 3 – Divisão dos movimentos *por intervalo de tempo* e *por célula espacial*: para que os movimentos em um dado intervalo se encaixem perfeitamente dentro dos limites de cada célula, novas observações são acrescentadas à trajetória nesse intervalo, nos pontos que intersectam as bordas espaciais e temporais das células. Isso é necessário para uma correta análise dos dados de trajetória. Por exemplo, na Figura 3(a) os movimentos no intervalo (00h13, 00h33) ultrapassam os limites espaciais das células c23 e c22. Para que os movimentos se encaixem dentro dessas células, acrescenta-se a observação 00h18 ao intervalo (00h13, 00h33) para distribuir seus movimentos entre (00h13, 00h18) e (00h18, 00h33), os quais respeitam os limites de c23 e c22, respectivamente. Esse é um exemplo da *divisão por célula espacial*. A *divisão por intervalo de tempo* é similar. Por exemplo, os movimentos no intervalo (01h53, 02h13) ultrapassam a barreira temporal, sendo assim, é acrescentada a observação 02h00 a esse intervalo. Considerando células espaço-temporais com duração de uma hora. É mostrada na Figura 3(c) a divisão dos movimentos da trajetória H1::C1.

Passo 4 – Sumarização dos movimentos por célula espaço-temporal: após a divisão dos movimentos por célula espaço-temporal (passos 3 e 4), os movimentos dentro de cada célula são sumarizados e armazenados na forma de um único registro no fato movimento. Na Figura 3(d) são ilustrados os movimentos da trajetória H1::C1, divididos e sumarizados por célula espaço-temporal.

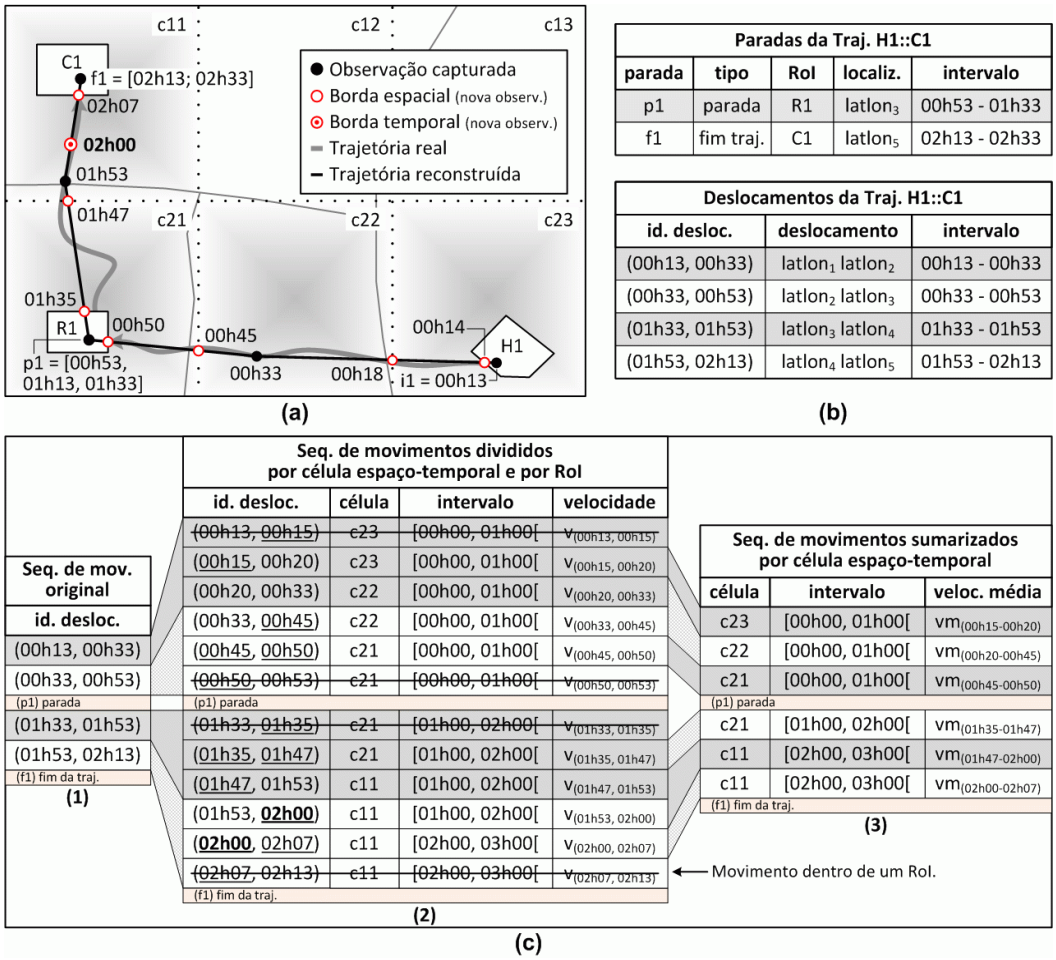
Para possibilitar múltiplas representações para trajetórias – isto é, roll-up para trajetórias – após a execução dos passos apresentados nos parágrafos acima, é necessário extrair e armazenar as direções do movimento, como descrito na seção a seguir.

5.2. Agregação por Direção dos Movimentos

Um dos desafios dos DWTrs é proporcionar múltiplas representações para trajetórias (Pelekis *et al.*, 2008), ou seja, proporcionar a representação de trajetórias e movimentos sobre diferentes perspectivas e níveis de granularidade. Por exemplo, para uma mesma trajetória, pode-se desejar visualizar o deslocamento do objeto móvel entre bairros, ou entre RoIs, ou de hora em hora, entre outras representações. Para resolver esse problema, a solução proposta consiste em obter múltiplas representações através de agregações das células espaço-temporais da trajetória (seu elemento mais básico), proporcionada pelo conjunto de dimensões direção do movimento.

Por exemplo, na Figura 3(a), é exibida a trajetória H1::C1, armazenada no *fato movimento* através da sequencia de células espaço-temporais [c23:00h, c22:00h, c21:00h, c21:01h, c11:01h, c11:02h]. Os *movimentos entre regiões adjacentes* de H1::C1 podem ser representados por *bairro* e por *RoI*, através das sequencias de movimentos [(*bairro1*, *bairro3*), (*bairro3*, *bairro4*), (*bairro4*, *bairro2*)] e [(H1, R1), (R1, C1)], res-

pectivamente, como mostrado na Figura 4(a). Para obter essas representações através do uso de agregações, cada célula de H1::C1 deve estar ligada (indicado por →) a um registro direção do movimento, como segue: (*ligação por bairro*) [c23:00h → (*bairro1*, *bairro3*)], [c22:00h, e c21:00h → (*bairro3*, *bairro4*)], [c21:01h, c11:01h, e c11:02h → (*bairro3*, *bairro4*)]; e (*ligação por RoI*) [c23:00h, c22:00h, e c21:00h → (H1, R1)], [c21:00h, c11:01h, c11:02h → (R1, C1)]. Sendo assim, para representar os movimentos de H1::C1 por bairro, basta selecionar a trajetória e colocar como critérios de agregação: o bairro-origem e o bairro-destino.



Paradas da Traj. H1::C1

parada	tipo	RoI	localiz.	intervalo
p1	parada	R1	$latlon_3$	00h53 - 01h33
f1	fim traj.	C1	$latlon_5$	02h13 - 02h33

Deslocamentos da Traj. H1::C1

id. desloc.	deslocamento	intervalo
(00h13, 00h33)	$latlon_1$ $latlon_2$	00h13 - 00h33
(00h33, 00h53)	$latlon_2$ $latlon_3$	00h33 - 00h53
(01h33, 01h53)	$latlon_3$ $latlon_4$	01h33 - 01h53
(01h53, 02h13)	$latlon_4$ $latlon_5$	01h53 - 02h13

Seq. de mov. original

id. desloc.
(00h13, 00h33)
(00h33, 00h53)
(p1) parada
(01h33, 01h53)
(01h53, 02h13)
(f1) fim da traj.

Seq. de movimentos divididos por célula espaço-temporal e por RoI

id. desloc.	célula	intervalo	velocidade
~~(00h13, 00h15)~~	~~c23~~	~~[00h00, 01h00[~~	~~$v_{(00h13, 00h15)}$~~
(00h15, 00h20)	c23	[00h00, 01h00[	$v_{(00h15, 00h20)}$
(00h20, 00h33)	c22	[00h00, 01h00[	$v_{(00h20, 00h33)}$
(00h33, 00h45)	c22	[00h00, 01h00[	$v_{(00h33, 00h45)}$
(00h45, 00h50)	c21	[00h00, 01h00[	$v_{(00h45, 00h50)}$
~~(00h50, 00h53)~~	~~c21~~	~~[00h00, 01h00[~~	~~$v_{(00h50, 00h53)}$~~
(p1) parada			
~~(01h33, 01h35)~~	~~c21~~	~~[01h00, 02h00[~~	~~$v_{(01h33, 01h35)}$~~
(01h35, 01h47)	c21	[01h00, 02h00[	$v_{(01h35, 01h47)}$
(01h47, 01h53)	c11	[01h00, 02h00[	$v_{(01h47, 01h53)}$
(01h53, **02h00**)	c11	[01h00, 02h00[	$v_{(01h53, 02h00)}$
(**02h00**, 02h07)	c11	[02h00, 03h00[	$v_{(02h00, 02h07)}$
~~(02h07, 02h13)~~	~~c11~~	~~[02h00, 03h00[~~	~~$v_{(02h07, 02h13)}$~~
(f1) fim da traj.			

Seq. de movimentos sumarizados por célula espaço-temporal

célula	intervalo	veloc. média
c23	[00h00, 01h00[	$vm_{(00h15\text{-}00h20)}$
c22	[00h00, 01h00[	$vm_{(00h20\text{-}00h45)}$
c21	[00h00, 01h00[	$vm_{(00h45\text{-}00h50)}$
(p1) parada		
c21	[01h00, 02h00[	$vm_{(01h35\text{-}01h47)}$
c11	[02h00, 03h00[	$vm_{(01h47\text{-}02h00)}$
c11	[02h00, 03h00[	$vm_{(02h00\text{-}02h07)}$
(f1) fim da traj.		

Figura 3. Carga de dados para a trajetória H1::C1: (a) representação de H1::C1 sobre o mapa; (b) observações capturadas de H1::C1 divididas entre paradas e movimentos; (c) divisão completa dos movimentos de H1::C1 por célula espaço-temporal em (2), e sua respectiva sumarização em (3).

Da forma apresentada, para cada representação desejada, é necessário manter uma dimensão direção do movimento, e uma chave estrangeira para relacionar o fato as dimensões, o que pode aumentar significativamente o volume do fato. Para reduzir o número de dimensões necessárias, uma forma adequada é unir essas dimensões [Kimball *et al.*, 2002]. Dessa forma, as células de H1::C1 passam a ser ligadas aos registros: [c23:00h → [(*bairro1*, *bairro3*), (H1, R1)]], [c22:00h, e c21:00h → [(*bairro3*, *bairro4*), (H1, R1)]], e [c21:01h, c11:01h, e c11:02h → [(*bairro4*, *bairro2*), (R1, C1)]], como mostrado na Figura 4(b). O inconveniente dessa solução, é que a união de dimensões pode gerar uma dimensão muito grande, como será discutido na Seção 6.2.

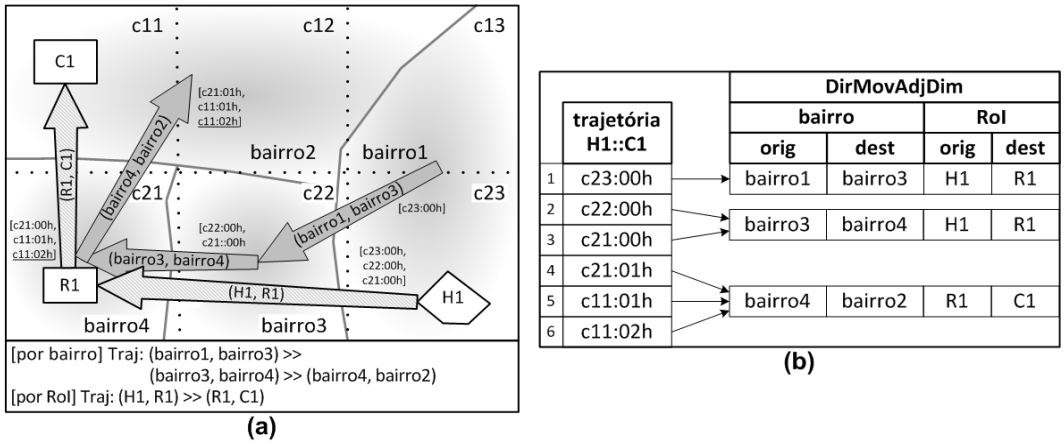


Figura 4. Representação por *direção do movimento entre regiões adjacentes*, para a trajetória H1::C1, em: (a) por bairro e por RoI. (c) Células de H1::C1 e sua ligação com a dimensão DirMovAdjDim.

6. Experimentos

Para avaliar o nível de compactação proporcionado pelo modelo proposto para dados de trajetória, diversos testes de carga de dados foram realizados sobre uma mesma base de trajetórias, mas usando diferentes configurações para cada carga executada. A configuração da base de dados é descrita na Seção 6.1, e o tamanho dos fatos e dimensões nas Seções 6.2 e 6.3, respectivamente.

6.1. Base de Dados de Trajetórias

Para realizar as consultas e os experimentos almejados, o DWTr foi povoado com uma base de dados de trajetórias gerada a partir de um *sintetizador de trajetórias semânticas* desenvolvido. Para dar um caráter mais realístico à base de dados, o protótipo desenvolvido permite criar problemas de tráfego a partir de um conjunto de configurações pré-estabelecidas definidas pelo usuário. Através dessas configurações é possível definir: quando e onde devem ocorrer congestionamentos; os locais com tráfego intenso; as regiões com maior ou menor velocidade; os locais de parada; o tempo de parada; os locais mais visitados pelos objetos móveis; entre outras características. Para dar um caráter ainda mais realista a base gerada, a rota de cada trajetória – isto é, a sequencia de coordenadas espaciais que vão da origem ao destino da trajetória – é obtida do Google Maps. De posse desses dados, o protótipo simula o deslocamento do objeto móvel.

Para gerar a base de dados sintética, simulou-se o comportamento de um conjunto de 2.000 objetos móveis, que se movimentaram na cidade de Aracaju/Sergipe, durante os meses de janeiro a junho de 2009. São aproximadamente 6.400 RoIs, e 1,8 milhão de trajetórias, uma média de 1.000 trajetórias por objeto móvel. No mundo real, isso equivale a cerca de 1 bilhão de observações, sendo 82 milhões referentes a *movimentos* e 931 milhões a *paradas*. Considerando observações capturadas a cada 20 segundos, e objetos móveis realizando em média 4 paradas de 2 horas por dia. Para tornar a leitura dos dados mais rápida e reduzir o espaço ocupado em disco, as observações referentes a paradas já são armazenadas na forma sumarizada. Dessa forma, consegue-se armazenar todas as *paradas* em 4 milhões de registros. A base de dados possui aproximadamente 15 GB de arquivos de texto no formato JSON[1].

[1] JSON: *JavaScript Object Notation*, é um formato similar ao XML, porém sua especificação é mais simples. É descrito pela RFC 4627. O site oficial: www.json.org. Acesso em: 02 fev 2011.

6.2. Tamanho das Tabelas de Fatos: Parada e Movimento

Para avaliar o nível de compactação proporcionado pelo modelo proposto para dados de trajetória em relação ao modelo clássico, duas baterias de testes foram realizadas: (i) *variando apenas o tamanho das células*, o intervalo de captura é sempre o mesmo, 20 segundos; e (ii) *variando apenas o intervalo de captura entre as observações*, o tamanho de cada célula foi fixado em 200 x 200 m^2. Os resultados dos experimentos (i) e (ii) são exibidos nas Tabelas 1(a) e 1(b), respectivamente.

Do **experimento (i)** é possível concluir que: (a) *a sumarização das paradas é a maior responsável pela compactação das trajetórias*, sua proporção de compactação (quantidade de dados compactada dividido pela quantidade de dados original) foi de 0,5%; e (b) para movimentos, como esperado, *quanto maior o tamanho das células, maior a taxa de compactação*, pois mais movimentos são sumarizados em uma mesma tupla.

Analisando os resultados do **experimento (ii)**, é possível perceber a grande quantidade de dados armazenada usando o modelo clássico, até mesmo para intervalos de captura longos (acima de 1 minuto), onde as trajetórias capturadas são imprecisas. Para facilitar a comparação dos modelos clássico e proposto, nesse experimento, a solução proposta não usa a interpolação linear local para reconstruir os movimentos das trajetórias. Ao invés disso, supõe-se um método capaz de recuperar sempre os mesmos movimentos da trajetória, independente do intervalo de captura usado. Dessa maneira, a forma espacial da trajetória reconstruída é sempre a mesma.

Embora, nos experimentos (i) para células menores que 200 x 200 m^2, e (ii) para intervalos de captura maiores que 20 segundos, a compactação dos movimentos gera mais dados que o número de observações. Isso ocorre porque o intervalo de captura usado é grande demais para o tamanho da célula, consequentemente, durante a carga de dados, novas observações são acrescentadas à trajetória para que seus movimentos se encaixem dentro dos limites das células, como discutido na Seção 5.2 (passo 3).

6.3. Tamanho das Dimensões Direção do Movimento

Para reduzir o número de dimensões *direção do movimento* e, consequentemente, o número de chaves estrangeiras no fato, a solução encontrada consiste em unir algumas dessas dimensões em uma só. Entretanto, essa solução possui um inconveniente: essas uniões podem gerar uma dimensão grande, o que pode levar a perda de desempenho. Para avaliar o tamanho dessas dimensões, vários experimentos foram realizados, envolvendo: (i) a *união das dimensões direção do movimento entre regiões adjacentes*; e (ii) a *união das dimensões direção do movimento* no tempo. Considerando células de tamanho 200x200 m^2, e a hierarquia de agregação 1x1 < 3x3 < 5x5 < 9x9 < 15x15, onde cada nível da hierarquia permite agregar um dado conjunto de células vizinhas. Por exemplo, cada agregação no nível 2 (3x3) agrega 9 células vizinhas da base do cubo (nível 1), gerando uma grade regular cujas células possuem 600x600 m^2 (isto é, 3 × 200x200 m^2). Os resultados dos experimentos (i) e (ii) são mostrados nas Tabelas 2(a) e 2(b), respectivamente.

Não foram encontrados problemas relacionados à união das dimensões no experimento (i), o número de registros gerados ficou abaixo de 1 milhão de tuplas, o que é recomendado para uma dimensão [Kimball *et al.*, 2002], mesmo quando se deseja analisar a movimentação entre pequenas regiões (por exemplo, regiões com 200x200 m^2). Entretanto, no experimento (ii), a união de duas ou mais dimensões ultrapassou o número de tuplas recomendado. Portanto, para analisar a direção do movimento entre regiões menores que 1x1 km^2, não é recomendada a união de dimensões.

Número de observações referentes a *movimentos* 82 milhões, *paradas* 935 milhões, *total* 1 bilhão.									
Tamanho da Célula			**50x50 m²**	**100x100 m²**	**200x200 m²**	**300x300 m²**	**1x1 km²**	**2x2 km²**	**3x3 km²**
Fato Movimento	**Núm. de Tuplas**	em milhões	262	132	66	44	14	7	5
	Proporção (em %)		318	160	80	54	17	9	7
Fato Parada	**Núm. de Tuplas**	em milhões	4						
	Proporção (em %)		0,5						
Paradas + Mov - Proporção (em %)			26,4	13,5	7	4,9	1,8	1,2	1

(a)

Intervalo de Captura		**10 seg.**	**20 seg.**	**1 min.**	**2 min.**	**3 min.**
Clássico: Mov/Paradas/Total	em milhões	165 / 1.871 / 2.037	82 / 935 / 1.018	27 / 311 / 339	13 / 155 / 169	9 / 103 / 113
Proposto: Mov/Paradas/Total		66 / 4 / 71				
Proporção (em %)		40,1 / 0,2 / 3,5	80,2 / 0,5 / 7	240,7 / 1,4 / 20,9	481,3 / 2,9 / 41,8	722 / 4,3 / 62,7

(b)

Tabela 1. Tamanho dos fatos, *parada* e *movimento*, usando o modelo proposto. Resultados dos experimentos: (a) variando apenas o tamanho das células da grade. (b) variando apenas o intervalo de captura entre observações.

Células	**Núm. de Reg.**
200x200 m² :: Níveis 1..5	485.900
200x200 m² :: Níveis 1..5 + Bairro	635.636
1x1 km² :: Níveis 1..5	28.170
1x1 km² :: Níveis 1..5 + Bairro	35.346
Bairro	1.324

(a)

Tamanho das Células	**20 min.**	**20, e 40 min.**	**20, 40, e 60 min.**
200x200 m²	535.730	1.697.040	2.588.365
1x1 km²	8.570	137.840	210.237

(b)

Tabela 2. Tamanho das dimensões sobre diferentes configurações. (a) Dimensões *Direção do Movimento entre Regiões Adjacentes* (DirMovAdjDim). (b) Dimensões *Direção do Movimento entre Regiões* (DirMovDim) no tempo.

7. Conclusões

Neste trabalho, é proposto um modelo para DW de Trajetórias (DWTr) que permite analisar o comportamento dos objetos móveis *sobre* e *entre* as regiões no espaço e tempo, o que é proporcionado pelo uso de *células espaço-temporais* e *dimensões direção do movimento* como critérios de agregação. Para amenizar o problema da grande quantidade dos dados de trajetória, propõe-se compactar trajetórias através da sumarização de suas paradas e movimentos. Com isso, conseguiu-se reduzir drasticamente o tamanho do fato *parada* e, de forma significativa, o tamanho do fato *movimento*, como mostrado através dos experimentos realizados. Para analisar o deslocamento dos objetos móveis entre as regiões, é mantido um conjunto de dimensões, que armazenam todas as direções do movimento (em termos de origem-destino) possíveis, para o conjunto de trajetórias armazenado. Como é mostrado através dos experimentos, a solução mostrou-se eficaz para a dimensão *direção do movimento entre regiões adjacentes* (DirMovAdjDim), o tamanho da dimensão permaneceu abaixo do valor máximo recomendado (1 milhão de tuplas), mesmo quando se analisa o deslocamento entre células pequenas (algumas dezenas de metros). Entretanto, o mesmo não aconteceu para a dimensão *direção do movimento entre regiões* no tempo (DirMovDim), a solução foi eficaz apenas para células grandes (acima de 1x1 km^2), quando se analisa o deslocamento entre células pequenas, o tamanho da dimensão ultrapassou o valor máximo recomendado. Para esses casos, é recomendado cautela para a união entre dimensões.

Como sugestões de trabalhos futuros, destacamos o desenvolvimento de: (i) operadores TrOLAP para agrupamento de trajetórias similares [Baltzer *et al.*, 2008]; (ii) métodos para reconstrução de trajetórias [Marketos *et al.*, 2008] e detecção de paradas [Bogorny *et al.*, 2009], visto que o monitoramento dos objetos móveis gera apenas da-

dos brutos, onde o início e fim das trajetórias ainda não são conhecidos, e não existe distinção entre paradas e movimentos, o que é fundamental para análise correta de trajetórias; (iii) métodos para enriquecer trajetórias com informações semânticas de forma automática; e (iv) métodos mais robustos para reconstrução dos movimentos da trajetória pois, embora a interpolação linear local seja um método simples e eficiente, não leva em consideração os dados sobre a infra-estrutura de rede sobre a qual os objetos móveis se movem (por exemplo, o mapa de ruas). Essas informações poderiam ser usadas para aproximar ainda mais as trajetórias reconstruídas das trajetórias reais.

Referências

Andrienko, G. e Andrienko, N. (2008). Spatio-Temporal Aggregation for Visual Analysis of Movements. *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pág. 51–58.

Andrienko, G., Andrienko, N., e Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *SIGKDD Explor. Newsl.*, 9(2):38–46.

Baltzer, O., Dehne, F., Hambrusch, S., e Rau-Chaplin, A. (2008). OLAP for trajectories. In *Database and Expert Systems Applications*, volume 5181 of *Lecture Notes in Computer Science*, pág. 340–347. Springer Berlin / Heidelberg.

Bédard, Y., Merrett, T., e Han, J. (2001). *Geographic Data Mining and Knowledge Discovery*, Capítulo: *Fundamentals of spatial data warehousing for geographic knowledge discovery*, pág. 53–73. CRC Press.

Bogorny, V., Kuijpers, B., e Alvares, L. O. (2009). ST-DMQL: A semantic trajectory data mining query language. *International Journal of Geographical Information Science*, 23(10):1245–1276.

Departamento Nacional de Trânsito (DENATRAN) / Fundação Getúlio Vargas (FGV) (2001). *Manual de Procedimentos para o Tratamento de Pólos Geradores de Tráfego.* Brasília – DF. Disponível em: http://www.denatran.gov.br/publicacoes/download/PolosGeradores.pdf. Acesso em: 02 fev 2011.

Gomez, L. I., Kuijpers, B., e Vaisman, A. A. (2008). Aggregation languages for moving object and places of interest. In *SAC'08: Proceedings of the 2008 ACM symposium on Applied computing*, pág. 857–862, New York, NY, USA. ACM.; Springer.

Kimball, R., Ross, M., e Merz, R. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley Computer Publishing, 2ª edição.

Kuijpers, B. e Vaisman, A. A. (2007). A data model for moving objects supporting aggregation. In *ICDE Workshops*, pág. 546–554.

Orlando, S., Orsini, R., Rafaeta, A., and Silvestri, A. R. C. (2007). *Trajectory data warehouses: Design and implementation issues.* Journal of Computing Science and Engineering (JCSE), 1(2):211–232.

Pelekis, N., Raffaeta, A., Damiani, M. L., Vangenot, C., Marketos, G., Frentzos, E., Ntoutsi, I., e Theodoridis, Y. (2008). Towards trajectory data warehouses. In *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*, Capítulo 7, pág. 189–211. Springer Publishing Company, Incorporated.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., e Vangenot, C. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1):126–146.