Um Modelo de Dados para Trajetórias de Objetos Móveis com suporte a Agregação de 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 the analysis of moving objects over and between regions in different levels of granularity. The model enables the segmentation of trajectories into components, such as stop and movement, carrying semantic information that assign meaning to the trajectory. To reduce the amount of data, the trajectories are stored compactly through the summarization of stops and movements.*

Resumo. *Neste trabalho é proposto um modelo conceitual para Data Warehouse de Trajetórias que permite analisar o comportamento dos objetos móveis sobre e entre regiões sobre diferentes níveis de granularidade. O modelo permite a segmentação de trajetórias em componentes, tais como parada e movimento, os quais podem transportar informações semânticas que dão significado à trajetória. Para amenizar o problema da grande quantidade de dados, as trajetórias são armazenadas de forma compactada através da sumarização de suas paradas e movimentos.*

1. Introdução

A popularização dos dispositivos móveis cientes de localização (*location-aware mobile devices*), tais como telefones celulares e GPSs (*Global Positioning System*), possibilitou o monitoramento em larga escala dos 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 desses dados permite descobrir padrões de comportamento que podem ser explorados por uma grande variedade de domínios. Por exemplo: (i) no gerenciamento de tráfego urbano, para identificar quais são as rotas mais utilizadas; (ii) no gerenciamento do transporte público, para melhorar a distribuição das linhas de ônibus; (iii) no estudo das trajetórias de turistas, para descobrir quais são locais mais visitados; e (iv) no estudo da migração de pássaros, para identificar rotas migratórias, entre outras.

Embora a análise de trajetória já faça parte do cotidiano de muitas aplicações, ela ainda é feita de forma rudimentar. Por exemplo, no gerenciamento de tráfego urbano, a análise dos dados de trajetória ainda é feita a partir dos dados sumarizados obtidos por sensores espalhados pela cidade, que capturam informações sobre a movimentação dos veículos em uma região, como por exemplo, a velocidade média. Entretanto, além desses dados não serem ricos em detalhes, eles são limitados ao que ocorre sobre uma região, logo, o deslocamento dos veículos entre as regiões não é capturado. Para uma melhor análise, esses dados poderiam ser substituídos por informações detalhadas sobre as trajetórias individuais de cada veículo, capturadas por GPSs. Isso permitiria descobrir, por exemplo, a origem-destino das trajetórias, as rotas comumente utilizadas, a velocidade em cada trecho da trajetória, as paradas realizadas, entre outras informações.

A base de dados gerada a partir do monitoramento dos objetos móveis é formada por um conjunto de dados brutos capturados por GPSs. Para transformar essa massa de dados bruta em informações úteis, uma forma adequada é disponibilizá-la em um *Data Warehouse* (DW) [Kimball *et al.*, 2002], um banco de dados otimizado para lidar com grandes volumes de dados de forma eficiente. Para aplicações com dados convencionais, DWs têm sido usados com sucesso no decorrer das últimas décadas. O principal objetivo de armazenar dados de trajetórias em um DW é permitir a análise de trajetórias no estilo OLAP [Leal *et al.*, 2011], como também, integrar diversas fontes de dados em uma só. A análise OLAP permite visualizar trajetórias sobre diferentes perspectivas e níveis de granularidade. Entretanto, a natureza dos dados de trajetória e a grande quantidade desses dados impõem certos desafios para a construção e manutenção do DW, dentre eles: **(i)** *o monitoramento dos objetos móveis gera apenas dados brutos* que, para a maioria das aplicações, não são suficientes para extrair informações úteis. Logo, 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 das trajetórias* [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 demasiadamente longo, impossibilitando a sua análise [Orlando *et al.*, 2007].

É proposto neste artigo um modelo semântico para Data Warehouse de Trajetórias (DWTrs) com suporte à agregação por direção dos movimentos. Pretende-se, com esse modelo **(i)** permitir analisar no estilo OLAP a movimentação dos objetos móveis sobre e entre as regiões no espaço e tempo, **análise orientada a tráfego** e **análise orientada a trajetória**, respectivamente. Em geral, os trabalhos na literatura proporcionam apenas *análise orientada a tráfego*, consequentemente, não permitem analisar a movimentação dos objetos *entre* as regiões, e visualizar trajetórias sobre diferentes níveis de granularidade, roll-up para trajetórias (maior contribuição do trabalho); **(ii)** possibilitar a **modelagem de trajetórias semânticas** em DWTrs, que permite dividir uma trajetória em diversos componentes semânticos, tais como parada e movimento, que podem transportar anotações que dão significado à trajetória [Spaccapietra *et al.*, 2008]. Diferente da maioria dos trabalhos analisados, que armazenam apenas dados brutos de trajetórias e se limitam a recuperar informações como a velocidade média ou o número de objetos móveis em uma região; **(iii)** para evitar **o problema da grande quantidade dos dados de trajetória** [Orlando *et al.*, 2007], os trabalhos relacionados que oferecem análise orientada a trajetória armazenam apenas a sua origem-destino, ignorando os movimentos intermediários, consequentemente, não podem oferecer análise orientada a tráfego. Para amenizar esse problema, e proporcionar análise orientada a tráfego e trajetória, propõe-se compactar trajetórias mediante a sumarização de suas paradas e movimentos. Dessa forma, consegue-se reduzir de forma significativa o tamanho dos fatos, como comprovado através dos experimentos realizados.

Este trabalho pressupõe que a massa de dados bruta sobre a movimentação dos objetos móveis já foi previamente processada e transformada em trajetórias semanticamente enriquecidas. Não são preocupações deste trabalho, questões tais como: (i) *receber o fluxo de dados contínuo* (*continuous data stream*) gerado pela captura das observações; (ii) *identificar quais são as trajetórias ali contidas*, e (iii) *enriquecê-las com informações semânticas*. Os trabalhos de Braz *et al.* (2007), Marketos *et al.* (2008), e Bogorny *et al.* (2009) tratam estes aspectos.

O restante deste artigo está organizado como segue: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve o cenário de aplicação usado durante os exem-

plos deste trabalho; a Seção 4 apresenta a forma adotada para representação de trajetórias; a Seção 5 descreve o modelo proposto e como proporcionar agregação por direção dos movimentos da trajetória; a Seção 6 apresenta a base de dados de trajetórias usada durante as consultas e experimentos; a Seção 7 apresenta algumas consultas exemplo, para mostrar o poder de expressividade do modelo proposto; a Seção 8 oferece uma discussão sobre os experimentos realizados para comprovar a redução no tamanho dos fatos, parada e movimento. Finalmente, a Seção 9 apresenta 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 Braz *et al.* (2007), Orlando *et al.* (2007), Marketos *et al.* (2008) e Baltzer *et al.* (2008) foram pioneiros nesse sentido: os três primeiros investigam como armazenar e agregar trajetórias armazenados na forma sumarizada, usando as tecnologias de DW tradicionais. Para isso, o espaço geográfico é dividido por uma grade regular formada por um conjunto de células. 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). Partindo de uma base de dados bruta sobre a localização espaço-temporal dos objetos, os autores investigam como extrair as trajetórias dessa base, transformar, e carregar esses dados. Enquanto, Baltzer *et al.* (2008) propõem um novo operador OLAP, *group_trajectories*, para agregação de trajetórias similares, que permite identificar quais objetos se movimentaram em paralelo durante um dado intervalo de tempo e/ou possuem rotas similares.

Segundo Andrienko e Andrienko (2008) as trajetórias de objetos móveis podem ser analisadas sobre dois pontos de vista: (i) visão orientada a tráfego e (ii) visão orientada a trajetórias, ou como prefere-se chamar neste trabalho, análise orientada a tráfego e análise orientada a trajetórias. Na *análise orientada a tráfego* o objetivo é analisar o comportamento dos objetos móveis *sobre* uma dada região em diferentes intervalos de tempo. Ela foi adotada por Braz *et al.* (2007), Marketos *et al*. (2008), e Orlando *et al.* (2007). Enquanto a *análise orientada a trajetórias* tem como objetivo analisar o deslocamento dos objetos móveis *entre* as regiões em termos de origem-destino do movimento. Ela é adotada por Baltzer *et al.* (2008), Gomez *et al.* (2008), Kuijpers e Vaisman (2007), e Spaccapietra *et al.* (2008). Dependendo do modelo adotado, ambos os tipos de análise estão disponíveis. Por simplificação, neste trabalho adota-se o termo *análise de trajetórias* para expressar ambos os tipos de análise. Mais formas de visualização e análise de trajetórias estão disponíveis em Andrienko e Andrienko (2008).

Data Warehouses Espaciais [Bédard *et al.*, 2001] são empregados por Kuijpers e Vaisman (2007), e Gomez *et al.* (2008) para analisar trajetórias. Segundo os 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) usa a mesma arquitetura de Kuijpers e Vaisman (2007), mas seu modelo distingue *paradas* de *movimentos* em trajetórias, além de compactar trajetórias através da sumarização de suas paradas e movimentos, representados na forma da transição entre paradas (por exemplo, da Casa C1 para o Trabalho Tr1).

Giannotti *et al.* (2007) usam técnicas de mineração de dados para descobrir quais são as regiões mais visitadas pelos objetos móveis em uma dada região. Enquanto Jensen (2002) apresenta alguns desafios enfrentados para construção de bancos de dados de objetos móveis, com ênfase no processamento de consultas que envolvem a loca-

lização corrente dos objetos móveis (por exemplo, *Nesse exato momento, qual é a localização de cada objeto móvel monitorado?*).

A solução proposta por Leal et al. (2011) é similar a deste trabalho. Os autores (i) permitem análise orientada a tráfego e trajetória; (ii) estendem o modelo de Spaccapietra et al. (2008); (iii) consideram que os componentes semânticos das trajetórias já são conhecidos; e (iv) utilizam um data warehouse. Entretanto, diferente deste trabalho, os autores armazenam todas as observações capturadas das trajetórias no DW, o que possui um inconveniente, a quantidade de dados armazenada pode ser muito grande, inviabilizando as consultas OLAP, como será visto na Tabela 1(a). Além disso, a solução não permite realizar a operação de roll-up para os movimentos da trajetória (maior contribuição deste trabalho).

Em geral, os trabalhos na literatura proporcionam apenas *análise orientada a tráfego*, e alguns destes conseguem resolver o problema da grande quantidade dos dados de trajetória, como Braz *et al.* (2007), Marketos *et al.* (2008) e Orlando *et al.* (2007). Entretanto, poucos trabalhos proporcionam *análise orientada a trajetórias*, tais como Baltzer *et al.* (2008), Gomez *et al.* (2008), Kuijpers e Vaisman (2007), e Spaccapietra *et al.* (2008), sendo que nenhum deles permitem analisar a direção dos movimentos das trajetórias no estilo OLAP. Além disso, dos trabalhos analisados, apenas Gomez *et al.* (2007) e Spaccapietra *et al.* (2008) distinguem paradas de movimentos, o que é fundamental para a análise correta de trajetórias, como será visto na Seção 4.

3. Cenário de Aplicação

Nesta seção é apresentado 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, ela 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, que captura sua localização espaço-temporal a cada 20 segundos e transmite, não necessariamente em tempo real, esses dados para um servidor, que processa esses dados e os transforma em trajetórias semanticamente enriquecidas.

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; **(ii)** locais mais frequentados e quando isso ocorre: local de trabalho, casa, bares; **(iii)** rotas comumente usadas para ir de um lugar ao outro; e **(iv)** meio de transporte utilizado. Além disso, são mantidas informações sobre cada cidade analisada e dados espaciais como: ruas (representadas por polilinhas), bairros (representados por polígonos) e regiões de interesse (ou RoIs - *Regions of Interest* - representados por polígonos, representando lugares como hotéis e restaurantes). 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 deve 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 através da *análise orientada a tráfego* e *análise orientada a trajetórias*, respectivamente.

4. Representação de Trajetórias

Para analisar trajetórias de forma adequada é necessário distinguir paradas de movimentos em trajetórias, e manter informações semânticas para dar significado às trajetórias. 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 (ou em outras palavras, por dois conjuntos de informações): a *face geométrica* e a *face semântica*.

A **face geométrica** é o conjunto de dados espaço-temporais brutos sobre a movimentação do objeto. Por simplificação, essa face 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 intermediários* da trajetória entre duas observações consecutivas, é usada a função de *interpolação linear local*, a qual considera que um objeto móvel se desloca em linha reta a uma velocidade constante entre duas observações, ver Figura 1(a).

A **face semântica** corresponde ao conjunto de informações que dão significado à trajetória, e podem ser usadas para dividir a *face geométrica* em diversos componentes semânticos, tais como parada e movimento, os quais podem transportar anotações que dão significado à parte da trajetória a qual pertence, como ilustrado na Figura 1(b).

A união da face geométrica com a face semântica tem como resultado, uma **trajetória espaço-temporal semântica**, onde cada observação (x y t) da face geométrica está associada a um componente da face semântica. Sendo assim, esse tipo de trajetória pode ser representado na forma: **[** $I:[(x_1\ y_1\ t_1)]$; $M_1:[(x_2\ y_2\ t_2), ..., (x_i\ y_i\ t_i)]$; $P_1:[(x_{i+1}\ y_{i+1}\ t_{i+1}), ..., (x_{i+j}\ y_{i+j}\ t_{i+j})]$ ...; $F:[(x_{m+1}\ y_{m+1}\ t_{m+1}), ..., (x_{m+n}\ y_{m+n}\ t_{m+n})]$ **]**, onde $i, j, m, n \in Z$, com $i < j < m < n$. Sendo que (i) I define a observação que marca o *início da trajetória*; (ii) M_1 delimita a sequencia de observações referentes a um *movimento*; (iii) P_1 delimita a sequência de observações referentes a uma *parada*; e (iv) F_1 referente a parada que corresponde ao *fim da trajetória*. Além disso, esses componentes transportam informações semânticas, como mostrado na Figura 1(c).

No modelo de Spaccapietra *et al.* (2008), o deslocamento de qualquer objeto móvel tem um objetivo, e é esse objetivo que define a origem-destino das trajetórias, não as paradas. Por exemplo, na Figura 1 o objeto móvel teve como objetivo sair de casa C1 para o trabalho Tr1, e foi esse objetivo que definiu a origem-destino da trajetória T2. As paradas $P2_1$ (parada no engarrafamento) e $P2_2$ (parada em um posto de gasolina) em T2 não tem relação com seu objetivo, podem ter acontecido por mera casualidade. Diferente de Spaccapietra *et al.* (2008), neste trabalho, todas as paradas correspondem a Regiões de Interesse (RoIs), e todas as trajetórias são delimitadas pelo fim de duas paradas, não necessariamente consecutivas (veja Figura 1, que T2 é delimitado por $[t_{fim\text{-}traj\text{-}f1}, t_{fim\text{-}traj\text{-}f2}]$, isto é, pelo fim das paradas F1 e F2).

Na análise de trajetórias é importante distinguir paradas de movimentos, pois a inclusão de dados sobre paradas na análise de movimentos pode provocar forte discrepância entre os valores analisados e os reais. Por exemplo, ao se incluir dados de paradas no cálculo da velocidade média de uma região, pode-se ter uma forte impressão de que a velocidade na região analisada está muito baixa, o que pode não ser verdade, devido à influência dos dados de paradas (por exemplo, de carros estacionados), cuja velocidade é zero ou muito próximo a isso. De forma similar, acontece com a análise de paradas ao se incluir dados de movimentos.

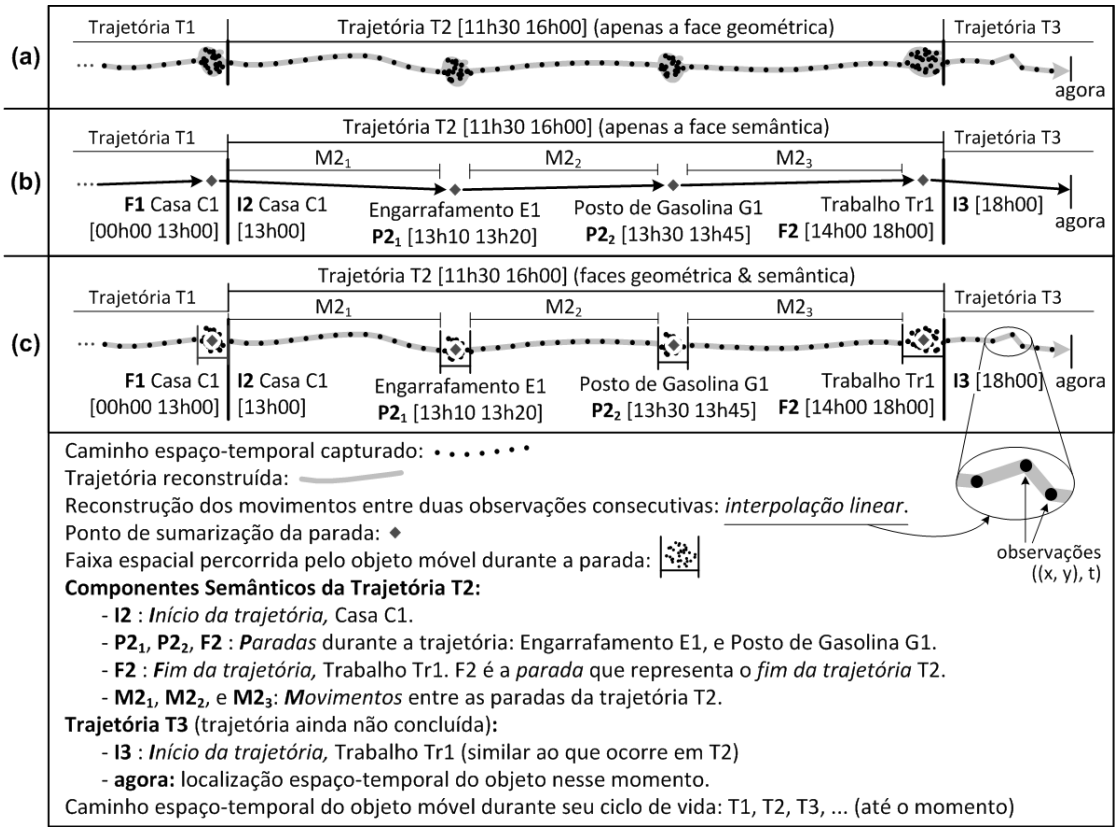


Figura 1. Diferentes representações para uma mesma trajetória. (a) apenas a face geométrica, dados espaço-temporais brutos; (b) apenas a face semântica, trajetória definida através de seus componentes; (c) trajetória espaço-temporal semântica, dados espaço-temporais & informações semânticas.

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. Para atingir os objetivos almejados, o modelo proposto incorpora as dimensões e fatos exibidos na Figura 2, os quais são descritos nos parágrafos a seguir. Os critérios de agregação e medidas apresentados aqui são algumas das sugestões possíveis.

Objeto Móvel (ObjMovDim): une as dimensões demográfica e tecnográfica. A dimensão *demográfica* mantém dados sobre os objetos móveis, como por exemplo, nome, sexo, idade, profissão e estado civil. A dimensão *tecnográfica* mantém dados sobre o dispositivo de localização usado, assim como a precisão do GPS usado.

Trajetória (TrajDim): dimensão descritiva, contém as informações sobre a trajetória como um todo. Basicamente possui informações: (i) *espaciais*: origem e destino da trajetória; (ii) *temporais*: início e fim da trajetória; e (iii) *descritivas*: objetivo da trajetória (por exemplo, saindo de casa para o trabalho), dados sobre o veículo utilizado na trajetória (atributo *veículo*).

Célula (CelulaDim): dimensão espacial, armazena as células espaciais da grade regular. Em geral, tem como hierarquia *celNivel1* < *celNivel2* < ... < *celNivelN* < *bairro*, onde cada nível da hierarquia permite agregar um dado número de células adjacentes na base do cubo (celNivel1). Por exemplo, considerando a hierarquia de

células – *200x200 m² (celNivel1 1x1) < 600x600 m² (celNivel2 3x3)* – cada célula do nível celNivel2 representa o resultado da agregação de 9 (3x3) células vizinhas do celNivel1, gerando assim uma grade regular cujas células possuem 600x600 m² (ou seja, (3*200)x(3*200) m²). Por isso, o tamanho da célula de cada nível é múltiplo do tamanho da célula base (celNivel1).

Tempo (TempoDim): dimensão temporal. Em geral, possui a hierarquia *min20 < hora < dia < mês < ano*, onde *min20* é um intervalo de tempo referente a 20 minutos. Essa dimensão também pode manter os eventos que ocorreram em cada intervalo de tempo (por exemplo, shows, partidas de futebol, acidentes de trânsito).

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 (o polígono que representa o RoI). Em geral, possui a hierarquia *roi < celNivel1 < ... < celNivelN < bairro*.

Direção do Movimento: representada pelas dimensões (i) *DirMovETDim*: mantém a Direção do Movimento entre as regiões Espaço-Temporais, através do par origem-destino do movimento e/ou de um tipo caractere (por exemplo, norte, sul, leste, oeste, ...), para um dado intervalo de tempo (por exemplo, do *bairro1* para o *bairro3* no intervalo de 20 min.); e (ii) *DirMovAdjDim*: mantém a Direção do Movimento entre regiões Adjacentes, similar a *DirMovETDim*, mas é independente dos intervalos de tempo.

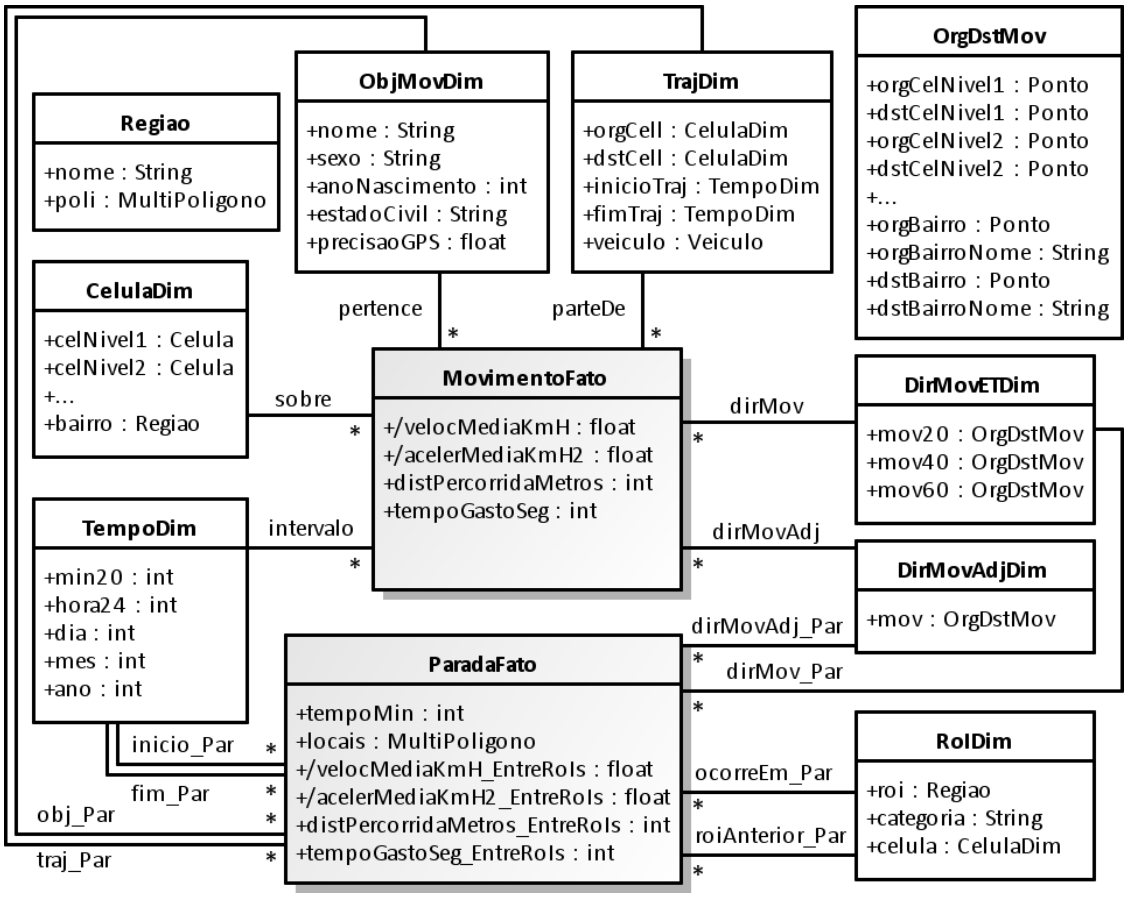


Figura 2. Diagrama UML do modelo proposto.

O modelo proposto possui dois fatos: (i) **fato parada** (*ParadaFato*) mantém os dados referentes às paradas na forma sumarizada por parada em um RoI representado basicamente pela medida *tempo de parada*, mas dependendo do nível de detalhe exigido

pode possuir a medida *locais de parada* (um dado espacial), um conjunto de polígonos que representam os locais de paradas específicos do objeto móvel dentro do RoI; e (ii) **fato movimento** (*MovimentoFato*) mantém os dados referentes aos movimentos de forma sumarizada por célula espaço-temporal. Representado pelas medidas *velocidade média*, *aceleração média*, *espaço percorrido*, *tempo gasto*, entre outras.

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: *DirMovETDim*, *DirMovAdjDim* e *RoIDim* (através de suas duas associações com *ParadaFato*, *roiAnterior_Parada* e *ocorreEm_Par*, representando a origem e destino do movimento entre RoIs, respectivamente). Na Seção 5.2 é descrito como funciona a agregação por direção dos movimentos.

Na Figura 2, na dimensão *TrajDim*, os atributos *orgCell, dstCell, inicioTraj*, e *fimTraj* são do tipo dimensão, dessa forma, podem ser interpretados como uma chave estrangeira para a dimensão de seu respectivo tipo (por exemplo, *orgCell*, célula de origem da trajetória, é uma chave estrangeira para dimensão CelulaDim). O mesmo ocorre na dimensão RoIDim com o atributo *celula*. Logo, trata-se de um *esquema floco de neve* (*snowflake schema*). Embora, por questões de desempenho, o esquema estrela seja o mais recomendado em DWs segundo Kimball *et al.* (2002), para esse caso específico, com muitos atributos espaciais, é preferível o floco de neve, porque permite acrescentar mais informações semânticas à dimensão com um melhor custo-benefício (desempenho-armazenamento). Mas, para comprovar está afirmação, seria necessário realizar alguns testes, o que foge ao escopo deste trabalho no momento.

5.1. Carga de Dados

Nesta seção, são descritos os passos necessários para transformar e sumarizar as observações da trajetória capturada de forma a se adequarem ao modelo proposto. Este trabalho pressupõe que os componentes da trajetória, tais como parada e movimento, já foram previamente identificados e as suas anotações semânticas já estão incluídas. Dessa forma, o algoritmo da carga de dados tem como entrada uma trajetória espaço-temporal semanticamente enriquecida, como descrito na Seção 4. Considera-se também que as trajetórias capturadas são precisas e feitas por GPSs em intervalos de tempo regulares e pequenos quando comparados com a velocidade do objeto (por exemplo, a cada 20 seg. para veículos em uma área urbana). Sendo assim, os passos para povoar o DWTr são:

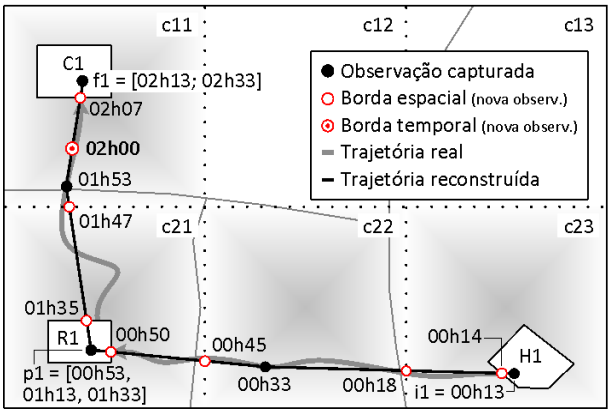
Passo 1 – Sumarização das observações referentes às paradas: as observações referentes a cada parada da trajetória são sumarizadas e armazenadas na forma de um único registro no fato parada. Sendo assim, na Figura 3(a) as observações no intervalo (00h53, 01h33) referentes à parada p1 são sumarizadas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

Passo 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 borda temporal (cada célula possui uma janela temporal de uma hora, nesse caso, o movimento começou na janela tempo de 01h00 as 02h00 e se estendeu até a janela de 02h00 as 03h00), sendo assim, é acrescentada a observação 02h00 a esse movimento. 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 (passo 3), os movimentos dentro de cada célula são sumarizados e armazenados na forma de um único registro no fato movimento. Na Figura 3(c) são ilustrados os movimentos da trajetória H1::C1, divididos e sumarizados por célula espaço-temporal.



(a)

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

Movimentos da Traj. H1::C1		
id. movimento	movimento	intervalo
(00h13, 00h33)	$latlon_1 \rightarrow latlon_2$	00h13 - 00h33
(00h33, 00h53)	$latlon_2 \rightarrow latlon_3$	00h33 - 00h53
(01h33, 01h53)	$latlon_3 \rightarrow latlon_4$	01h33 - 01h53
(01h53, 02h13)	$latlon_4 \rightarrow latlon_5$	01h53 - 02h13

(b)

Seq. de mov. original

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

(1)

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.			

(2)

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.		

(3)

← Movimento dentro de um RoI.

(c)

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).

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

5.2. Construindo as Dimensões Direção do Movimento

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, permitir visualizar o deslocamento do objeto móvel entre bairros, ou entre RoIs, ou de hora em hora por bairro, 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 (dimensões DirMov).

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)], respectivamente, 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.

Da forma apresentada, para cada representação desejada é necessário manter uma dimensão DirMov, e uma chave estrangeira para relacionar o fato às dimensões, o que pode aumentar significativamente o volume do fato. Para reduzir o número de dimensões necessárias, uma forma adequada é unir essas dimensões. 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).

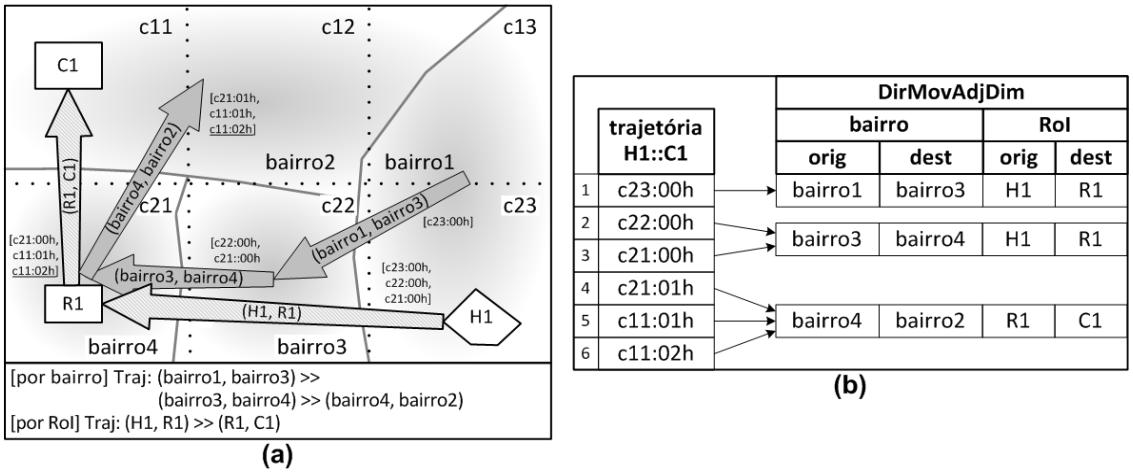


Figura 4. A trajetória H1::C1 representa de duas formas diferentes. Em (a) movimentos da trajetória por bairros entre regiões adjacentes; e em (b) as células da trajetória ligadas à dimensão que proporciona essas representações.

O inconveniente dessa solução, é que a união de muitas dimensões (por exemplo, 18 dimensões) pode ter como resultado uma dimensão muito grande (acima de 1 milhão de tuplas), gerando perda de desempenho [Kimball *et al*., 2002]. Uma solução simples para amenizar o problema é: (i) evitar a união de muitas dimensões em uma só, mantendo duas ou mais dimensões união; e (ii) unir dimensões de forma otimizada, buscando combinar as dimensões que minimizem o número de tuplas distintas, reduzindo assim o tamanho da dimensão união. A melhor forma de realizar essas combinações, ainda não foi estudada. No entanto, partindo dos experimentos realizados, já se sabe que mesmo unindo muitas dimensões DirMov de forma não otimizada, o volume de dados total do DW (fatos + dimensões DirMov) é bem menor do que o caso que não considera a união das dimensões, como será discutido na Seção 8.2 (ver Tabela 4).

6. 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. Com essas 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 à base gerada, a rota de cada trajetória (ou seja, a sequência de coordenadas espaciais que vão da origem ao destino da trajetória) é obtida a partir do *Google Maps*. De posse desses dados, o protótipo simula o deslocamento do objeto móvel. O mapa da cidade e os bairros – isto é, os polígonos – foram obtidos da *Wikimapia* (similar ao Google Maps, mas permite obter dados espaciais de regiões).

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, estado de 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 de maneira 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 (*JavaScript Object Notation*, um formato similar ao XML).

7. Consultas Exemplo

Para mostrar o poder de expressividade do modelo proposto são apresentadas algumas consultas, as quais buscam atender aos requisitos necessários para gerenciamento de tráfego urbano, descritos na Seção 3. As consultas exemplo envolvem *análise orientada a tráfego* e *análise orientada a trajetórias*, como será discutido a seguir.

7.1. Análise Orientada a Tráfego

A análise orientada a tráfego tem como objetivo analisar situações de tráfego, ou seja, analisar o comportamento dos objetos móveis *sobre* uma dada região espacial em diferentes intervalos de tempo [Andrieko e Andrieko, 2008]. Esse tipo de análise envolve apenas *agregação por região*.

Agregação por Região

Para exemplificar a agregação das medidas por região, considere a **(Consulta 1)**: *Qual a velocidade média dos objetos móveis circulando na cidade no dia 15-fev-2009 por região da cidade e por hora?* Considere uma região da cidade como uma célula do nível 1 (200x200 m^2). Essa consulta é resolvida através do fato *Movimento* da seguinte forma:

Sobre o fato Movimento (MovimentoFato): (i) seleção dos atributos *célula do nível 1* (celNivel1), *hora* (hh24), e *velocidade* (velocidade_kmh), usando a função de agregação média para a medida *velocidade*; (ii) restrição temporal para a data 15-fev-2009; e (iii) agrupamento por *célula do nível 1* e por *hora*.

Na Figura 5(a) é ilustrada a consulta em SQL, e na Figura 5(b) é mostrado o resultado da consulta. Os quadrados na cor verde indicam regiões onde os objetos fluem tranquilamente, 60 km/h ou mais; em azul, regiões com velocidade média entre [30, 60[km/h; em vermelho, regiões problemáticas, onde os objetos circulam abaixo de 30 km/h, indicando possível ocorrência de congestionamentos. O tamanho dos quadrados é proporcional ao valor do atributo analisado. Para facilitar a visualização dos resultados, considerou-se apenas os bairros *Getúlio Vargas* e adjacentes, no intervalo de tempo das 10h00 às 11h00 horas. Na Figura 5(c) é mostrado o resultado de uma consulta similar, que considera o número de paradas em RoIs por célula espaço-temporal do nível 1, onde os quadrados em vermelho representam as regiões que somadas representam 50% das paradas ou mais, e os quadrados azuis as demais regiões.

```
SELECT cd.cel_nivel_1 AS celula, td.hh24 AS hora,
       AVG( mf.velocidade_kmh ) AS veloc_media
  FROM MovimentoFato AS mf RIGHT JOIN
       CellDim       AS cd ON (mf.cell_id  = cd.cell_id ) INNER JOIN
       TempoDim      AS td ON (mf.tempo_id = td.tempo_id)
 WHERE td.data = to_date('15-fev-2009', 'DD-mon-YYYY')
 GROUP BY celula, hora
```

(a)

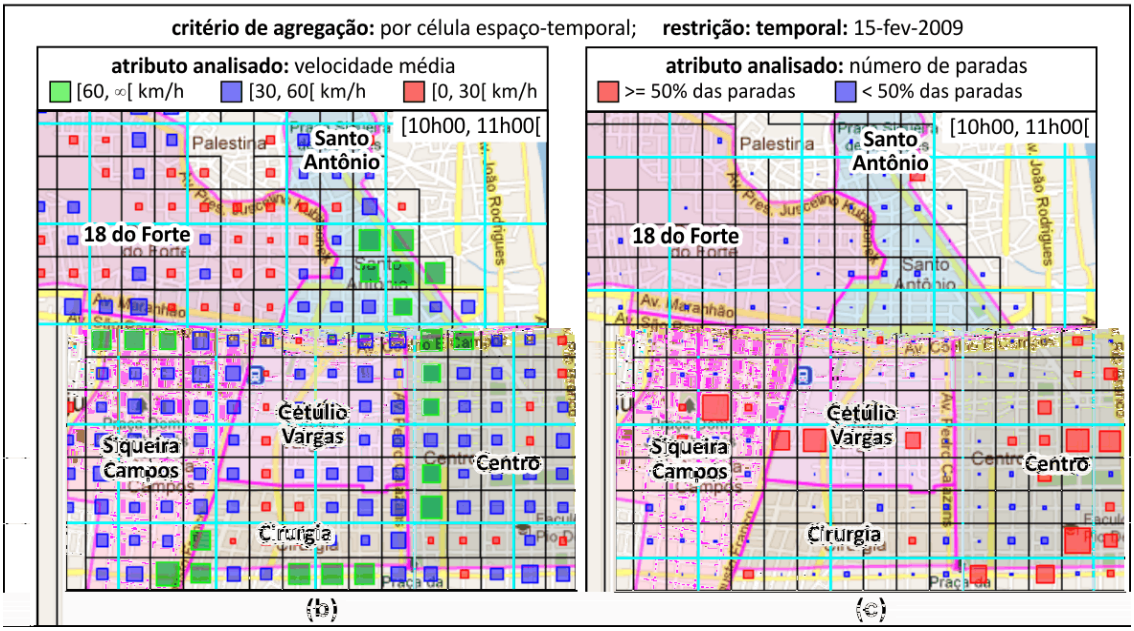


Figura 5. Agregação por Região: (a) consulta em SQL referente à (b); (b) velocidade média dos objetos por célula do nível 1; e (c) uma consulta similar, mas em termos do número de paradas em RoIs por célula.

Considerando o modelo ilustrado na Figura 2, as consultas podem usar qualquer uma das medidas: do (i) fato **movimento**, tais como velocidade, aceleração, distância percorrida, tempo gasto, e atributos derivados (por exemplo, número de objetos); ou do (ii) fato **parada**: tempo de parada, e atributos derivados (por exemplo, número de objetos parados); entre outras possibilidades. Agregadas através das hierarquias: (a) **espacial** celNivel1 (200x200 m^2) < celNivel2 (600x600 m^2) < celNivel3 (1x1 km^2) < celNivel4 (1.8x1.8 km^2) < celNivel5 (3x3 km^2) < bairro; (b) **temporal** 20 min < 40 min < 1 hora < dia < mês < ano. Para exemplificar os níveis de detalhe da grade adotada, na Figura 5(b) duas grades são visíveis, a preta com células do nível 1 (200x200 m^2) original da consulta, e na cor ciano com células do nível 2 (600x600 m^2), mas na Figura 6(b) na cor ciano é ilustrado células do nível 3 (1x1 km^2).

7.2. Análise Orientada a Trajetórias

Tem como objetivo analisar o deslocamento dos objetos móveis *entre* as regiões espaciais em diferentes intervalos de tempo [Andrieko e Andrieko, 2008]. As consultas nesse tipo de análise envolvem agregação por direção do movimento entre regiões. Para facilitar sua compreensão elas foram divididas em três categorias envolvendo: (i) agregação dos movimentos entre regiões; (ii) agregação por origem-destino; e (iii) agregação dos movimentos das trajetórias.

Agregação dos Movimentos entre Regiões

Permite visualizar o deslocamento dos objetos móveis *entre* as regiões no espaço em diferentes intervalos de tempo, independente da trajetória a qual pertencem. Para exemplificar esse tipo de agregação considere a **(Consulta 2):** *Qual o número de objetos móveis circulando na cidade no dia 15-fev-2009 entre bairros por hora?* Considere a direção do movimento entre regiões adjacentes. Essa consulta pode ser resolvida usando o fato movimento, como segue:

Sobre o fato Movimento (MovimentoFato): (i) seleção dos atributos *bairro de origem do movimento* (org_bairro), *bairro de destino do movimento* (dst_bairro), *hora* (hh24), e *número de objetos móveis* (obtido usando a função de agregação contagem distinta sobre o identificador do objeto móvel (obj_id)); (ii) restrição temporal para a data 15-fev-2009; e (iii) agrupamento por *origem do movimento*, *destino do movimento*, e *hora*.

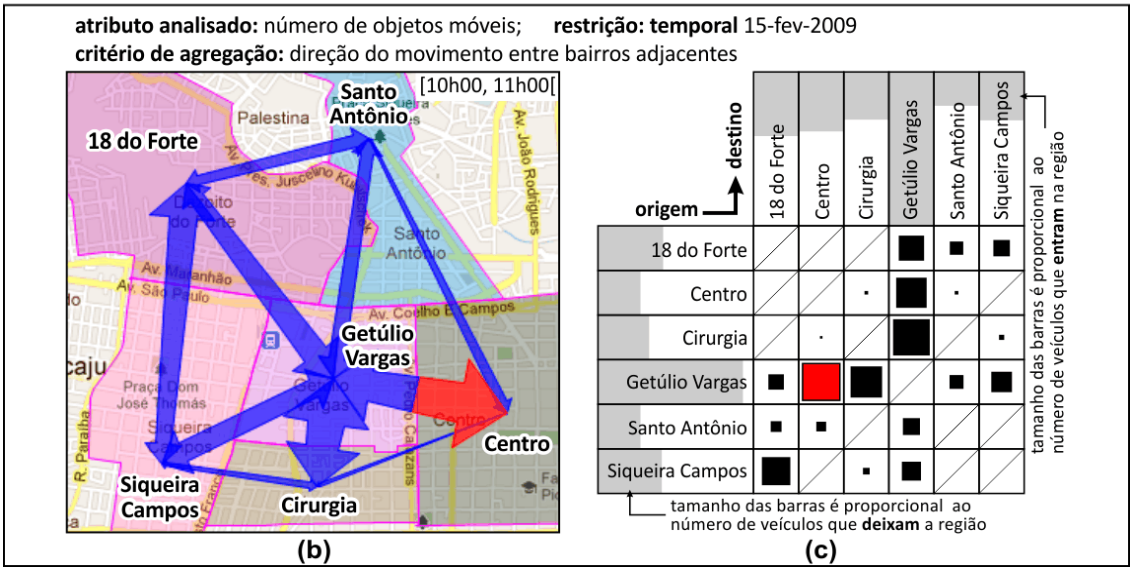
Na Figura 6(a) é ilustrada a consulta em SQL, e na Figura 6(b) é mostrado o deslocamento dos objetos móveis entre bairros adjacentes. O tamanho das setas é proporcional ao número de objetos, a seta em vermelho indica o deslocamento com maior número de objetos. Na Figura 6(c) é mostrado o mesmo resultado, mas na forma de uma matriz *origem* x *destino*, onde os quadrados são proporcionais ao número de objetos. As barras horizontais na coluna *origem* representam o número de objetos que *deixam* o respectivo bairro, e as barras verticais na linha *destino* denotam o número de objetos que *entram* no bairro. Como na Consulta 1, considerou-se apenas os bairros *Getúlio Vargas* e adjacentes.

De acordo com o esquema de dados usado, para a Consulta 2 ainda seria possível ver os movimentos das trajetórias: **(i)** por *direção do movimento entre regiões adjacentes* (DirMovAdjDim) entre: as células dos níveis de 1 à 5, e bairros; ou **(ii)** por *direção do movimento entre regiões espaço-temporais* (DirMovETDim) entre regiões a cada 20, 40, ou 60 minutos, onde essas regiões são células dos níveis de 1 a 5, e bairros. A principal diferença entre os itens (i) e (ii) é que apenas em (i) é possível ver todas as regiões por onde o objeto móvel passou. Por exemplo, supondo que um objeto saiu do

bairro 18 do Forte com destino ao bairro Centro realizando a trajetória [(18 do Forte, 10h00), (Siqueira Campos, 10h10), (Cirurgia, 10h15), (Centro, 10h35)], considerando (i) a *agregação por direção do movimento entre regiões adjacentes*, a trajetória seria representada através dos movimentos [(18 do Forte, Siqueira Campos), (Siqueira Campos, Cirurgia), (Cirurgia, Centro)], mas considerando (ii) *agregação por direção do movimento entre regiões espaço-temporais* por bairro a cada 20 minutos, têm-se [((18 do Forte, Cirurgia), 10h00-10h20), ((Cirurgia, Centro), 10h20-10h40)], o movimento sobre o bairro Siqueira Campos não é mostrado.

```
SELECT dmad.org_bairro AS origem, dmad.dst_bairro AS destino,
       td.hh24 AS hora, COUNT( DISTINCT mf.obj_id ) AS num_objmov
  FROM MovimentoFato AS mf   INNER JOIN
       DirMovAdjDim  AS dmad ON (mf.movadj_id = dmad.movadj_id) INNER JOIN
       TempoDim      AS td   ON (mf.tempo_id  = td.tempo_id   )
 WHERE td.data = to_date('15-fev-2009', 'DD-mon-YYYY')
 GROUP BY origem, destino, hora
```

(a)



(b) (c)

Figura 6. Agregação dos Movimentos entre Regiões. Em (a) consulta em SQL; (b) resultado da consulta sobre o mapa, e (c) resultado na forma de uma matriz origem x destino.

Agregação por Origem-Destino

Leva em consideração apenas a origem e destino da trajetória. Pode ser muito útil para analisar o fluxo (ou seja, o número) de objetos móveis entre as regiões, ou o tempo gasto pelos objetos no percurso. Para exemplificar esse tipo de agregação considere a **Consulta 3:** *Saindo de sua origem, quanto tempo os objetos móveis levam em média para atingir o centro da cidade nas horas de rush?* Considere que os objetos saem de um bairro origem e têm como destino o bairro Centro, e que se deseja descobrir a origem dos objetos móveis que levam mais tempo para atingir o centro.

Por simplificação não é mostrado uma figura dedicada a esse tipo de agregação. Mas nas Figuras 7(b) e 7(c) a seta na cor ciano é o resultado de uma agregação por origem-destino, onde todas as trajetórias dos objetos móveis que partem de uma célula do bairro 18 do Forte (origem) com destino a uma célula do bairro Centro são agregadas.

Agregação dos Movimentos da Trajetória

É similar a agregação por origem-destino, mas permite ver o deslocamento dos objetos móveis de região em região, desde a origem até o destino da trajetória, de forma detalhada. A agregação dos movimentos da trajetória pode ser útil para descobrir quais são os trechos mais problemáticos da trajetória de sua origem até seu destino.

Para exemplificar esse tipo de agregação considere a **Consulta 4:** *Qual é o tempo gasto pelos objetos móveis saindo do bairro 18 do Forte com destino ao bairro Centro, para cada parte das trajetórias?* Considere que uma parte da trajetória corresponde à direção do movimento entre células do nível 1. O código SQL dessa consulta é ilustrado na Figura 7(a). Suponha que o usuário fornece as trajetórias (*18 do Forte*, *Centro*). Para facilitar a visualização do resultado, suponha também que todas as trajetórias (*18 do Forte*, *Centro*) possuem a mesma rota (ver Figura 7(b)). As setas em vermelho indicam as regiões problemáticas, onde os objetos gastam mais de 50% do tempo para realizar o percurso (*18 do Forte*, *Centro*). Na Figura 7(c), é mostrado o resultado de uma consulta bem similar, mas a agregação considera a direção do movimento entre células do nível 2.

8. 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. O tamanho das bases de fatos e de dimensões é descrito nas Seções 8.1 e 8.2, respectivamente.

```
SELECT dmad.org_cel_nivel_1 AS org_move, dmad.dst_cel_nivel_1 AS dst_move,
       AVG(mf.tempo_gasto) AS tempo_gasto
  FROM MovimentoFato AS mf   INNER JOIN
       DirMovAdjDim  AS dmad ON (mf.movadj_id = dmad.movadj_id) INNER JOIN
 WHERE ( mf.obj_id   IN ([ids dos objetos]) ) AND -- obj_id & num_traj
       ( mf.num_traj IN ([trajetórias]) )         --  igual ao id da traj
 GROUP BY org_move, dst_move
```

(a)

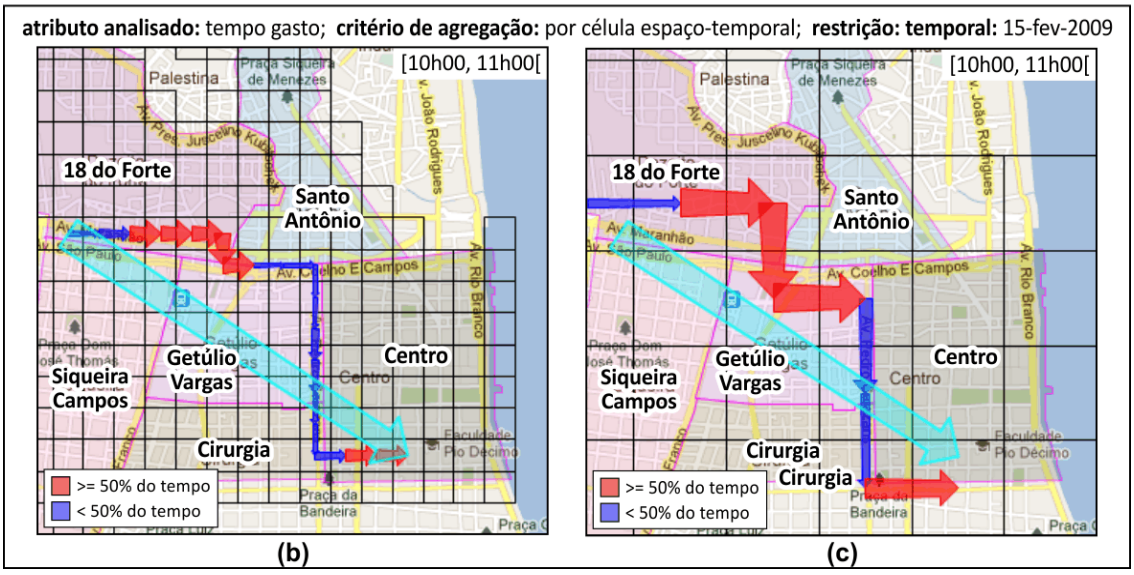


Figura 7. Agregação dos Movimentos da Trajetória. Em (a) código em SQL referente à (b); (b) trajetórias agregadas por direção do movimento entre regiões adjacentes; (c) mesma agregação, mas em um nível maior de granularidade.

8.1. Tamanho dos 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 das observações é sempre o mesmo, 20 segundos; e (ii) *variando apenas o intervalo de captura*, o tamanho de cada célula foi fixado em 200x200 m^2. Os resultados desses experimentos são exibidos nas Tabelas 1(a) e 1(b), e na Tabela 2 é mostrado o tamanho dos fatos em megabytes para alguns dos testes realizados no experimento (i).

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 sumarizada / quantidade de dados original) foi de 0,5%; e (b) para movimentos, como já era esperado, *quanto maior o tamanho das células, maior a 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.

Note-se que o experimento (i) para células menores que 200x200 m^2, e (ii) para intervalos de captura maiores que 20 segundos, a compactação dos movimentos tenha gerado mais dados que o número de observações originais. Isso ocorre porque o algoritmo da carga de dados reconstrói os movimentos intermediários da trajetória entre cada par de observações, para aproximá-la um pouco mais da real, o que pode acrescentar novas observações à trajetória capturada. Consequentemente, o número de observações da trajetória reconstruída pode ser maior do que o número de observações originais.

Experimento (i)		Considerando observações capturadas a cada 20 segundos							
Tamanho das células ➡		**50x50 m^2**	**100x100 m^2**	**200x200 m^2**	**300x300 m^2**	**1x1 km^2**	**2x2 km^2**	**3x3 km^2**	
Clássico	**Movimento**	82 milhões							núm. de tuplas em milhões
	Parada	1.017 milhões (≈ 1 bilhão)							
Proposto	**Movimento**	263	132	66,5	45	14	8	5,5	
	Parada	4,5 milhões							
Taxa de Compactação taxa = proposto / clássico	**Movimento**	318%	160%	80%	54%	17%	9%	7%	
	Parada	0,5%							

(a)

Experimento (ii)						
Intervalo de captura das observações ➡		**10 seg.**	**20 seg.**	**1 min.**	**2 min.**	**3 min.**
Clássico	**Movimento**	165	82	27	13	9
	Parada	1.871	935	311	155	103
Proposto	**Movimento**	66 milhões				
	Parada	4 milhões				
Taxa de Compactação taxa = proposto / clássico	**Movimento**	40%	80,2%	240%	481%	722%
	Parada	0,2%	0,5%	1,5%	3%	4%

(b)

Tabela 1. Número de tuplas dos fatos Parada e Movimento. Resultados do: (a) experimento (i), variando apenas o tamanho das células da grade; e (b) experimento (ii), variando apenas o intervalo de captura das observações.

De posse da Tabela 1, o tamanho dos fatos parada e movimento – isto é, o espaço de armazenamento ocupado por eles – foi calculado para três condições distintas: **caso 1**, *ignorando as dimensões direção do movimento*, ou seja, ignorando as suas chaves estrangeiras; e no **caso 2** considerando essas dimensões, mas primeiro, **caso 2a** *sem realizar as uniões*, e depois **caso 2b,** *realizando as uniões*. Os resultados dessas contas são mostrados na Tabela 2. Observa-se que sem realizar a união das dimensões (caso 2a) o tamanho dos fatos mais que quadruplicou (4,4 vezes maior que o caso 1), e realizando a união das dimensões (caso 2b), o tamanho dos fatos é 1,2 vezes maior que o caso 1. Portanto, analisando sobre esse ponto de vista, a união das dimensões vale a pena, mas ainda resta saber se o ganho obtido com a redução no tamanho dos fatos compensa o aumento no tamanho das dimensões, como será visto em detalhes na Seção 8.2.

Considerou-se nos casos 2a e 2b que as dimensões DirMov: no **caso 2a**, acrescentam 24 chaves estrangeiras (FKs) as tuplas de cada um dos fatos, são + 6 FKs das dimensões DirMovAdjDim + 18 FKs das dimensões DirMovETDim (considerando os intervalos de tempo 20 40 e 60 min); e no **caso 2b**, acrescentam 2 chaves estrangeiras, + 1 FK de DirMovAdjDim + 1 FK de DirMovETDim. Considerou-se uma FK como um inteiro de 4 bytes. Os fatos possuem os atributos definidos a seguir.

Tamanho das Células ➡	**200x200 m²**			**1x1 km²**			
	Caso 1	**Caso 2a**	**Caso 2b**	**Caso 1**	**Caso 2a**	**Caso 2b**	**TC***
Fato **Movimento**	1.777 MB	7.869 MB	2.284 MB	375 MB	1.663 MB	483 MB	29%
Fato **Parada**	154 MB	565 MB	188 MB	154 MB	565 MB	188 MB	33%
Movimento + Parada	1.931 MB	8.434MB	2.472 MB	529 MB	2.228MB	671 MB	30%

Resultados obtidos com base nos testes do experimento (i), ver Tabela 1.
Considerando 24 representações possíveis para as trajetórias. Para isso, o *caso 2a* requer +24 chaves estrangeiras (FKs) (18 DirMovETDim + 6 DirMovAdjDim), e *caso 2b* +2 FKs apenas (1 DirMovETDim + 1 DirMovAdjDim).
*** TC** = Taxa de Compactação (Caso 2b / Caso 2a), quanto menor o valor melhor.

Tabela 2. Tamanho dos fatos Parada e Movimento

O ***fato movimento*** possui no **caso 1**, tuplas de 28 bytes, provenientes das FKs *objId, trajId, celulaId, tempoId*, e das medidas *distância percorrida,* e *tempo gasto*, como ilustrado na Seção 5 Figura 2. Todos os atributos são inteiros de 4 bytes, com exceção de *tempoId* que ocupa 8 bytes. As medidas velocidade e aceleração não foram incluídas porque podem ser calculadas a partir das outras medidas; no **caso 2a**, o fato movimento possui tuplas de 124 bytes; e no **caso 2b**, tuplas de 36 bytes.

O ***fato parada*** possui no **caso 1**, tuplas de 36 bytes, provenientes das FKs *objId, trajId, celulaId, tempoInicioId, tempoFimId, roiId, roiAnteriorId*, e da medida *tempo de parada*. Todos os atributos são inteiros de 4 bytes, com exceção de *tempoInicioId* e tempoFimId que ocupam 8 bytes cada; no **caso 2a**, o fato parada possui tuplas de 132 bytes; e no **caso 2b**, tuplas de 44 bytes.

8.2. Tamanho das Dimensões Direção do Movimento

Para reduzir o número de dimensões *direção do movimento* (DirMov) e, consequentemente, o número de chaves estrangeiras no fato, a solução encontrada consiste em unir algumas dessas dimensões em uma só. O inconveniente dessa solução é que a união de muitas dimensões pode gerar uma dimensão muito grande, comprometendo o desempenho. Para avaliar o tamanho dessas dimensões, diversos experimentos foram realizados, envolvendo (1) a *união das dimensões Direção do Movimento entre regiões Adjacentes (DirMovAdjDim)*; e (2) a *união das Dimensões Direção do Movimento entre Regiões Espaço-Temporais (DirMovETDim)*.

Para os experimentos considerou-se uma hierarquia espacial com 6 níveis, para permitir visualizar a direção dos movimentos desde células até bairros, e até 3 intervalos de tempo (20 40 e 60 min). Sendo a hierarquia espacial: celNivel1 1x1 < celNivel2 3x3 < celNivel3 5x5 < celNivel4 9x9 < celNivel5 15x15 < bairro, onde no experimento (1) as células base possuem 200x200 m^2; e no experimento (2) células base possuem 1x1 km^2. Sendo assim, são necessários: no *experimento (1)* 6 dimensões DirMovAdjDim, independente do número de intervalos de tempo; e no *experimento (2)* 6 dimensões DirMovETDim para cada intervalo de tempo analisado.

Analisando os experimentos conclui-se: (i) não houve problemas na união das 6 dimensões DirMovAdjDim, mesmo quando a análise de movimentos era precisa (envolvendo desde pequenas regiões 200x200 m^2 até bairro); entretanto, (ii) quando se analisava os movimentos espaço-temporais de forma precisa para 2 ou mais intervalos de tempo (12 dimensões ou mais), a união das dimensões DirMovETDim apresentou problemas, o número de tuplas ultrapassou o máximo recomendado (1 milhão [Kimball *et al.* 2002]). Portanto, acima de 6 dimensões já é recomendado manter mais de uma dimensão união, para evitar dimensões muito grandes. Vale ressaltar que, mesmo no pior caso, o tamanho total do DW (fatos + dimensões DirMov) foi 2 vezes menor do que o caso que *não considera a união das dimensões* (caso 2a), como mostrado na Tabela 4. No caso 1, os fatos não possuem chaves estrangeiras para as dimensões DirMov. Para a Tabela 4 suponha os mesmos atributos da Tabela 2, mostrada na Seção 8.1.

DirMovAdjDim – Experimento (1)

Tamanho das células	R	Núm. de tuplas	Tam. da tupla	Tam. da dimensão
200x200 m^2	6	635 mil	316 Bytes	192 MB
1x1 km^2	6	35 mil		11 MB

R = Número de Represent. proporcionadas pela dim.

(a)

DirMovETDim – Experimento (2)

	Intervalos de tempo analisados	R	Núm. de tuplas	Tam. da tupla	Tam. da dimensão
200x200 m^2	**20 min**	6	535 mil	316 Bytes	161 MB
	20 e 40 min	12	1.697 mil	628 Bytes	1.016 MB
	20 40 e 60 min	18	2.588 mil	940 Bytes	2.320 MB
1x1 km^2	**20 min**	6	8 mil	316 Bytes	3 MB
	20 e 40 min	12	137 mil	628 Bytes	83 MB
	20 40 e 60 min	18	210 mil	940 Bytes	188 MB

R = Número de Representações proporcionadas pela dimensão.

(b)

Tabela 3. Tamanho das dimensões direção do movimento sobre diferentes configurações. Em (a) dimensões Direção dos Mov. Adjacentes; e (b) dimensões Direção dos Mov. entre as Regiões Espaço-Temporais.

Tamanho das Células ➡	200x200 m^2					1x1 km^2				
	Caso 1	R	Caso 2a	Caso 2b	TC	Caso 1	R	Caso 2a	Caso 2b	TC
Tamanho dos fatos + as dimensões DirMov	1.931 MB	12	5.182 MB	2.284 MB	44%	529 MB	12	1.379 MB	543 MB	39%
		18	6.803 MB	3.139 MB	46%		18	1.803 MB	623 MB	34%
		24	8.434 MB	4.443 MB	52%		24	2.228 MB	728 MB	32%

R = Número de Representações proporcionadas pelas dimensões DirMovETDim & DirMovAdjDim.
TC = Taxa de Compactação (caso 2b / caso 2a), quanto menor o valor melhor.
52% é o pior resultado, e 39% é o melhor resultado.

Tabela 4. Tamanho total do DWTr, fatos + as dimensões direção do movimento. No (*caso 1*) ignorando essas dimensões; e no (*caso 2*) as considerando, mas (*caso 2a*) sem realizar suas uniões, e (*caso 2b*) realizando suas uniões.

Para facilitar o cálculo do tamanho das tuplas das dimensões DirMov suponha que são formadas por atributos do tipo OrgDstMov (origem e destino do movimento). Esse tipo representa o conjunto de atributos necessários para visualizar os movimentos da trajetória para cada um dos 6 níveis de representação definidos na hierarquia espacial do parágrafo anterior. Para isso, OrgDstMov possui os atributos [(orgCelNivel1, dst-

CelNivel1), ..., (orgCelNivel5, dstCelNivel5), (orgBairro, dstBairro), (orgBairroNome, dstBairroNome)], como ilustrado na Seção 5 Figura 2. Todos esses atributos são do tipo Ponto (uma coordenada espacial), com exceção de orgBairroNome e dstBairroNome, que são Strings. O tipo Ponto ocupa 21 bytes, o mesmo espaço ocupado pelo tipo POINT no SGBD PostgreSQL, e os nomes dos bairros ocupam no máximo 30 bytes, o que equivale a 30 caracteres. Sendo assim, o tipo OrgDstMov ocupa 312 bytes, provenientes de: 6 pontos para a origem do movimento + 6 pontos de destino + nome do bairro de origem + nome do bairro de destino.

Representando as dimensões direção do movimento dessa forma, a dimensão DirMovAdjDim requer apenas 1 atributo do tipo OrgDstMov, independente do número de intervalos de tempo analisados. Sendo assim, cada uma de suas tuplas ocupa 316 bytes, 4 bytes da chave primária + 312 do tipo OrgDstMov. O mesmo não ocorre na dimensão DirMovETDim, a qual requer um atributo do tipo OrgDstMov para cada intervalo de tempo analisado. Sendo assim, para analisar os movimentos a cada 20, 40 e 60 minutos, são necessários 3 atributos, um para cada intervalo de tempo, gerando uma tupla de 940 bytes (4 da chave primária + 3*312 dos tipos OrgDstMov). Como DirMovETDim, em geral, possui mais atributos que a dimensão DirMovAdjDim, é de se esperar que suas uniões gere um conjunto de valores bem maior do que na união das dimensões DirMovAdjDim, como é possível perceber analisando a Tabela 3.

9. Conclusões

Neste trabalho, propõe-se 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 analisar o deslocamento dos objetos entre as regiões, é mantido um conjunto de dimensões, com todas as possíveis direções do movimento (em termos de origem-destino) para o conjunto de trajetórias armazenado. 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 discutido na Seção 8.1.

Para proporcionar análise orientada a trajetórias, os fatos estão associados a dimensões direção do movimento (DirMov), uma dimensão para cada representação desejada. Logo, quanto maior for o número de dimensões, maior será o tamanho dos fatos. Para contornar esse problema, as dimensões DirMov são unidas (ver Seção 5.2). Entretanto, a união de muitas dimensões pode gerar uma dimensão muito grande, comprometendo o desempenho. Para saber até quantas dimensões podem ser unidas, diversos experimentos foram realizados (ver Seção 8.2), considerando a hierarquia espacial com 6 níveis, e até 3 intervalos de tempo (20, 40 e 60 min), o que implica em: (i) 6 dimensões DirMovAdjDim, para os movimentos entre as regiões adjacentes; e (ii) 6 dimensões DirMovETDim para cada intervalo de tempo analisado, para os movimentos entre regiões espaço-temporais. Dos experimentos se conclui: (a) a união das dimensões DirMovAdjDim, não apresentou problemas, o número de tuplas permaneceu dentro dos limites aceitáveis (1 milhão [Kimball *et al.* 2002]); entretanto, (b) para representar os movimentos espaço-temporais para 2 ou mais intervalos de tempo (12 dimensões ou mais), o número de tuplas ultrapassou o aceitável. Para esses casos, é recomendado cautela na união das dimensões. Vale ressaltar, que mesmo no pior caso (unindo muitas dimensões), o volume de dados total do DW (fatos + dimensões DirMov) foi 2 vezes menor, do que o caso que não considera uniões.

Como sugestões de trabalhos futuros, destaca-se 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 dados 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.

Uma versão mais resumida deste artigo foi publicado no VII Simpósio Brasileiro de Sistemas de Informação (SBSI'11), ver [Almeida *et al.*, 2011].

Referências

Almeida C. A. S., Pires C. E., Schiel U. (2011). Data Warehouse de Trajetórias: um Modelo com Suporte à Agregação por Direção dos Movimentos. In *VII Simpósio Brasileiro de Sistemas de Informação* (SBSI'11), pp. 57-68. Salvador, Brasil.

Andrienko G. e Andrienko N. (2008). Spatio-temporal aggregation for visual analysis of movements. *IEEE Symposium on Visual Analytics Science and Technology* (VAST'08), pp. 51-58. Columbus, Ohio, USA.

Baltzer O., Dehne F., Hambrusch S., e Rau-Chaplin A. (2008). OLAP for trajectories. *Database and Expert Systems Applications*, Volume 5181 of *Lecture Notes in Computer Science*, pp. 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*, pp. 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.

Braz F., Orlando S., Orsini R., Raffaeta A., Roncato A., e Silvestri C. (2007). Approximate aggregations in trajectory data warehouses. Em ICDE *Workshops*, pp. 536–545.

Departamento Nacional de Trânsito (DENATRAN) / Fundação Getúlio Vargas (FGV) (2001). *Manual de procedimentos para o tratamento de pólos geradores de tráfego*, Brasília, DF, Brasil.

Giannotti, F., Nanni, M., Pinelli, F., e Pedreschi, D. (2007). Trajectory pattern mining. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pp. 330–339, New York, NY, USA. ACM.

Gomez L. I., Kuijpers B., e Vaisman A. A. (2008). Aggregation languages for moving object and places of interest. *Proceedings of the ACM symposium on Applied computing*, pp. 857-862, New York, NY, USA.

Jensen C. S. (2002). Research challenges in location-enabled m-services. Em *Mobile Data Management*, pp. 3–7.

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. *ICDE Workshops*, pp. 546-554, Istanbul, Turkey.

Leal B. C., Macêdo J. A. F., Times V. C., Casanova M. A., Vidal V. M. P., Carvalho M. T. M. (2011). From Conceptual Modeling to Logical Representation of Trajectories in DBMS-OR and DW Systems. *Journal of Information and Data Management*, Vol. 2, No. 3, October 2011, pp. 463–478.

Marketos G., Pelekis N., Frentzos E., Raffaeta A., Ntoutsi I., e Theodoridis, Y. (2008). Building real-world trajectory warehouses. Proc. *7th International ACM SIGMOD Workshop on Data Engineering for Wireless and Mobile Access (MobiDE' 08)*.

Orlando S., Orsini R., Rafaeta A., e 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*, Cap. 7, pp. 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.