

# Identificação de Estruturas Sociais em Registros Reais de Mobilidade Humana e Veicular

Danielle L. Ferreira<sup>1</sup>, Bruno A. A. Nunes, Carlos A. V. Campos<sup>1</sup>

<sup>1</sup>Programa de Pós-Graduação em Informática (PPGI)  
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)  
CEP 22290-240 – Rio de Janeiro – RJ – Brasil

bastuto@gmail.com, {danielle.ferreira,beto}@uniriotec.br

**Abstract.** *This work presents a methodology for extracting relevant information from human and vehicular mobility traces, allowing the characterization of user mobility, and their geographical preferences. Moreover, a data mining methodology is proposed to extract and distinguish community structures based on mobility and geographical preferences of mobile users. Such methodology allows the identification of communities without the need for traces on homophily and/or real social relationships between mobile entities. These traces are usually hard to get access to. The proposed methodology was applied to two real mobility datasets (human and vehicular mobility), used in the community identification process. Principal components were extracted from the features matrix generated after a number of transformations on the raw dataset. These allowed to make evident the dissimilarities amongst the different members of the identified communities via the proposed methodology. Finally, we show that the identified communities present in fact, statistically different attributes and distinct geographical preferences.*

**Resumo.** *O presente trabalho apresenta uma metodologia para extração de informações relevantes em registros de movimentação humana e veicular, permitindo caracterizar a mobilidade dos usuários, bem como a preferência geográfica dos mesmos. Além disso, é proposta uma metodologia de mineração de dados capaz de identificar e distinguir estruturas de comunidades baseada na mobilidade e nas preferências geográficas dos usuários. Tal metodologia permite que se identifique comunidades sem a necessidade de conjuntos de dados de informação sobre homofilia e/ou relacionamentos reais entre as entidades móveis. Este tipo de dados são normalmente menos acessíveis. A metodologia proposta foi aplicada a dois conjuntos de dados reais (movimentação humana e veicular), utilizados na identificação das comunidades. Para isso, foi realizada a extração de componentes principais da matriz de atributos gerada após a realização de transformações nos dados. Isso permitiu evidenciar as dissimilaridades entre os diversos membros das comunidades identificadas através da metodologia proposta. Por fim, mostra-se que as diversas comunidades identificadas possuem de fato, atributos estatisticamente distintos e preferências geográficas diferentes.*

## 1. Introdução

Redes oportunistas são uma interessante evolução e um importante paradigma em redes móveis ad-hoc. Aplicações baseadas em contatos oportunistas tornam-se atraentes principalmente na ausência de uma infraestrutura fixa de rede para comunicação. Tais

aplicações foram alavancadas pelo contínuo crescimento da popularidade de dispositivos móveis, observado nos últimos anos e pelos desafios apresentados em cenários específicos, tais como áreas rurais e cenários urbanos muito populosos. Neste contexto, contatos temporários e ocasionais entre usuários e seus dispositivos, apresentam-se como oportunidades de transmissão de dados, enquanto a mobilidade destes usuários pode ser vista como o meio de transporte destes dados [Han et al. 2012].

A disseminação de informação em redes oportunistas depende de redundância (i.e., replicação de mensagens) para reduzir a latência e aumentar a taxa de entrega das mensagens. Normalmente, o encaminhamento das mensagens é realizado através do envio de múltiplas cópias das mesmas encaminhadas para outros dispositivos na rede. Dessa forma, alcança-se baixa latência e alta taxa de entrega. Porém, devido às restrições de tamanho de buffer, largura de banda e durabilidade de bateria dos dispositivos móveis, se faz necessário encontrar um equilíbrio entre latência, taxa de entrega e replicação das mensagens. Portanto, para realizar um encaminhamento de mensagens de forma eficiente, é necessário um processo de encaminhamento adequado, buscando enviar as mensagens para dispositivos/indivíduos específicos de forma a maximizar as chances de entregar a mensagem com sucesso. Para tal, conhecer as estruturas sociais existentes entre os nós da rede pode ser determinante na frequência com a qual os dispositivos se encontram o que, por sua vez, impacta na eficiência do mecanismo de encaminhamento de mensagens.

Diversos trabalhos em redes oportunistas procuram descobrir relações ou comportamentos semelhantes entre os usuários que compõem a rede, de forma a utilizar estas informações na tomada de decisão de quando e para quem encaminhar mensagens [Li and Wu 2009, Chuah and Coman 2009, Hui et al. 2007]. Porém, a maioria dos esquemas de detecção de comunidades extraem informações das estruturas de comunidades através do tempo entre contatos dos usuários [Eagle and (Sandy) Pentland 2006] (isto é, contato entre pares), ou utilizando informações obtidas através de redes sociais ou de operadoras de redes de telefonia [Phithakkitnukoon et al. 2012, Motani et al. 2005]. Portanto, seria bastante interessante se fosse possível identificar comunidades através apenas do comportamento de mobilidade humana, ao invés de depender de informações obtidas de fatores externos.

Neste contexto, o objetivo desse trabalho é descobrir, através de ferramentas de mineração de dados, padrões de movimentação semelhantes, dentro de um grupo de usuários de forma a classificá-los em diferentes subgrupos. Chamamos esses subgrupos de comunidades. Para isso, foi proposto um método de extração de comunidades, baseado apenas em informações extraídas da mobilidade humana, através do qual é possível detectar eficientemente a estrutura de comunidades. O método está baseado no tratamento de dados, de forma a identificar e evidenciar as dissimilaridades entre os padrões de movimentação, dos usuários móveis. Assim, os registros de mobilidade (i.e. *traces*) foram inicialmente pré-processados para que se pudesse extrair medidas de interesse para identificação dessas comunidades, tais como preferência geográfica, tempo de pausa e velocidade média do usuário.

Em resumo, as contribuições do presente trabalho podem ser relacionadas da seguinte forma: A proposta de uma metodologia (1) para a detecção de estruturas de comunidades, através do uso de registros reais de mobilidade humana e/ou veicular; (2) para a extração da preferência geográfica do usuário, através do uso de registros de mobilidade

humana e/ou veicular, permitindo identificar diferentes padrões, dependendo do universo temporal definido (e.g. diferentes dias da semana); (3) para ser aplicada a um número grande de tipos de registros de mobilidade, ou seja, alta heterogeneidade dos dados, uma vez que a metodologia é genérica o suficiente. Esses registros podem ser facilmente coletados através de smartphone (posicionamento através de GPS, conectividade com diversos pontos de acesso WiFi, etc.), o que elimina a dependência de dados, muitas vezes difíceis ou custosos de serem obtidos, como os provenientes de operadoras de telefonia móvel, redes sociais online ou dados de homofilia (identificar entre diversos atores, semelhanças de diversas ordens, tais como: idade, sexo, naturalidade, profissão, nível de educação, renda, etc.). Além disso, apresenta (4) uma métrica de similaridade baseada nas características intrínsecas do movimento de mobilidade humana, independente de agentes ou informações externas, (5) um estudo quantitativo comparando três métodos distintos de agrupamentos.

A estrutura do restante deste documento está como se segue. A Seção 2 apresenta os trabalhos relacionados. A Seção 3, introduz brevemente os métodos de agrupamento clássicos encontrados na literatura. Na Seção 4 a metodologia proposta para identificação de comunidades em dados heterogêneos é apresentada. A metodologia proposta é validada em dois estudos de caso aplicados a traces reais de mobilidade na Seção 5. Por fim, na Seção 6, são apresentadas as considerações finais.

## **2. Trabalhos Relacionados**

Alguns trabalhos da literatura utilizam mineração de dados para detectar interesses em determinadas áreas geográficas pelos usuários. Em [Khetarpaul et al. 2011], os autores propõem um método para analisar a localização agregada de GPS dos usuários e extrair os interesses de localização dos usuários e ranque-los. Em [Zheng et al. 2009] os autores também usam trajetórias de GPS de usuários para minerar interesses de localização e sequências de viagem. Os autores em [Giannotti et al. 2007] mineram sequências similares de trajetórias de usuários para encontrar padrões de trajetórias e regiões de interesse, aplicando diferentes métodos para extração de padrões. Porém, esses trabalhos não consideram as características sociais dos usuários e, portanto, não tratam do problema do agrupamento de usuários.

Eagle e outros [Eagle and (Sandy) Pentland 2006] buscam reconhecer padrões sociais na atividade diária de usuários utilizando traces gerados de dispositivos móveis, de 100 usuários usando Bluetooth. Esse trabalho explora o comportamento do usuário e seu perfil de mobilidade para propor uma metodologia de identificação de comunidades baseado nas similaridades encontradas entre os diferentes usuários. Porém, o trabalho utiliza dados de encontros de Bluetooth e a localização do usuário é inferida pela localização das torres de celular, perdendo dessa forma a granularidade da movimentação dos nós móveis.

Os autores em [Ferrari et al. 2011] extraem padrões de redes sociais baseadas em localização dos usuários na cidade de Nova Iorque, utilizando a aplicação Twitter. Em [Tang et al. 2012] propõe um método para extração de similaridades entre usuários de diferentes redes sociais para agrupá-los em comunidades. Porém, as redes sociais fornecem informações sobre a localização do usuário ou seus interesses com grande granularidade, uma vez que as informações apenas são gravadas quando o usuário efetivamente utiliza a rede social. Por exemplo, posta imagens no Instagram, ou faz *chek-in* utilizando o Foursquare de posições geográficas específicas o que dificulta a obtenção de dados que possam

levar a um conjunto de preferências geográficas.

Alguns trabalhos propõem esquemas para extrair as relações sociais entre usuários e também de suas comunidades sociais [Girvan and Newman 2002, Newman 2004]. Identificar comunidades sociais facilita o encaminhamento de pacotes em redes oportunistas. Hui e outros [Hui et al. 2007] propuseram um esquema de detecção distribuído para *Pocket Switched Networks*, onde cada dispositivo sente e detecta sua própria comunidade analisando o histórico de dispositivos móveis que o mesmo encontrou. Apenas os eventos de encontro são usados para construir a relação social entre eles. Esses trabalhos utilizam dados obtidos de traces de redes oportunistas o qual contém apenas informações dos encontros entre os dispositivos móveis.

### 3. Algoritmos de Agrupamento para Identificação de Comunidades

Redes sociais humanas são conhecidas por apresentarem fortes características de agrupamento [Newman 2004]. Por exemplo, as pessoas passam mais tempo com família, amigos e colegas de trabalho, do que com desconhecidos. Portanto, é interessante encontrar componentes nos padrões de movimentação de usuários de dispositivos móveis, tais como velocidade, tempo de pausa e localização geográfica, de forma a representar essas características de agrupamento observadas no mundo real. Existe uma grande variedade de algoritmos de agrupamento que podem ser utilizados para esse fim [Jain et al. 1999, Duda et al. 2001]. Nesse trabalho aplicamos três algoritmos de agrupamento, amplamente utilizados na literatura, buscando responder a questão que surge sobre o algoritmo mais adequado para capturar as propriedades da movimentação do usuário:

**Agrupamento Hierárquico** - Esta técnica utiliza uma métrica de similaridade  $x_{i,j}$  entre pares de nós  $(i, j)$ , que busca avaliar o quão similares são os atributos comuns a um grupo de nós, tais como: interesses, localização, velocidade de movimentação, etc. Essa métrica vem da observação de que indivíduos preferem se relacionar com outros indivíduos que possuem interesses similares e/ou realizam ações similares. A cada passo do algoritmo de agrupamento, uma matriz contendo as métricas de distância entre os agrupamentos é calculada. Depois de criada esta matriz de similaridades, deve-se encontrar o menor valor da matriz de similaridades que identifica os dois agrupamentos mais similares entre si. Estes dois, então, são reagrupados, formando assim um novo agrupamento. Logo em seguida, a matriz de similaridades é recalculada, contendo agora um agrupamento a menos. Esse procedimento é refeito até que reste apenas um único agrupamento.

Neste trabalho, foi utilizado o critério de variância mínima, o qual minimiza a variância total dentro do do agrupamento. Ou seja, a cada passo encontra-se o par de grupos que leva ao menor aumento da variância total do cluster após sua união. Além disso, vale ressaltar que o método *Ward* tende a combinar clusters que tem um pequeno número de observações, além de alocar clusters que são do mesmo tamanho e esféricos.

**Agrupamento *k-Means*** - Essa categoria consiste em técnicas que otimizam algum critério, para particionar as observações em um número pré-determinado de grupos ( $k$ ), onde a soma dos quadrados dentro de cada grupo é minimizada. O algoritmo para agrupamento *k-means* opera da seguinte forma: (1) inicialmente o número  $k$  de grupos é especificado *a priori*, então os centróides iniciais são determinados (a escolha do centróide pode ser feita de forma aleatória, ou pode ser previamente especificada). Após essa

etapa, (2) as distâncias euclidianas entre cada observação e cada centróide do cluster são calculadas, e cada observação é atribuída ao cluster mais próximo. Na próxima etapa, (3) o centróide é novamente calculado usando as observações agrupadas no passo anterior. Os passos do algoritmo são repetidos até que não aconteçam mais trocas de membros dos grupos ou até não haver mais alterações nos centróides. Nesse método de agrupamento a escala das variáveis pode afetar esse critério, portanto é importante realizar a normalização dos dados. Cabe comentar que métodos de cluster que minimizam tal critério, tendem a produzir agrupamentos que tem forma de hiper-elipsoide.

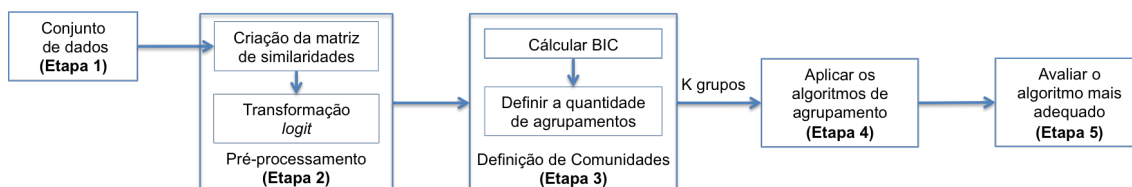
**Model-Based Cluster (MBC)** - MBC assim como os algoritmos anteriores também tem o objetivo de dividir os dados em grupos, porém o mesmo tenta resolver problemas encontrados nos algoritmos hierárquicos e k-means, que tendem a produzir grupos esféricos e de mesmo tamanho. Essa técnica utiliza o agrupamento hierárquico em uma de suas etapas para otimizar a classificação da função de similaridade. O MBC estima uma função de densidade de probabilidade de mistura finita [Dasgupta and Raftery 1995]. A abordagem de mistura finita assume que a função densidade de probabilidade pode ser modelada como a soma das densidades ponderadas dos componentes (ou grupos). Nessa função de densidade, cada componente representa um agrupamento e é possível utilizar a probabilidade a posteriori, que uma observação pertence ao componente de densidade, para atribuir o rótulo do grupo à observação. O MBC consiste em três etapas principais: (1) - Inicialmente é necessário especificar o número de densidades de componentes (ou grupos). Para tal, utiliza-se o agrupamento hierárquico aglomerativo para obter as partições iniciais dos dados. (2) - Aplicação do algoritmo EM (do inglês, *Expectation-Maximization*), no qual é baseado em uma estimativa de probabilidade máxima, utilizado para estimar a verossimilhança dos parâmetros da mistura. (3) - Por fim, uma vez que o algoritmo MBC assume que a mistura finita é composta de densidades normais multivariadas, onde restrições nas matrizes de covariância dos componentes levam a diferentes modelos, a técnica BIC (*Bayesian information criterion*) é utilizada para escolher o melhor modelo.

A próxima seção apresenta uma metodologia, que inclui a aplicação dos algoritmos de agrupamento apresentados acima, à dois conjuntos heterogêneos de dados reais de mobilidade humana, com o objetivo de tentar identificar similaridades entre os diversos indivíduos e agrupá-los adequadamente.

#### 4. Metodologia Proposta

O foco deste trabalho está na extração de estruturas sociais baseada apenas nas preferências geográficas e nos perfis de movimentação dos indivíduos, utilizando registros de movimentação reais. Dessa forma, esta seção apresenta a metodologia proposta de forma resumida, buscando atingir tal objetivo. As etapas do processo do processamento dos registros de movimentação até a definição das comunidades são mostradas na Figura 1. A metodologia opera como segue: (Etapa-1) - Os registros de mobilidade, inicialmente dados brutos, são tratados de forma a extrair os atributos desejados e construir a matriz de atributos. No presente trabalho essa matriz contém o tempo que cada usuário permanece em cada célula, seu tempo médio de pausa e sua velocidade média. Outros atributos podem ser utilizados em diferentes análises e a matriz não está limitada a estes apenas. (Etapa-2) - Os valores dos atributos são então normalizados e a transformação *logit* é aplicada para evidenciar as dissimilaridades entre os mesmos. (Etapa-3) - Aplicar

o BIC à matriz de atributos tratada na Etapa-2 e encontrar o número de agrupamentos  $k$  que maximiza o BIC. (Etapa-4) - Aplicar algoritmos de agrupamento e identificar comunidades. (Etapa-5) - Por fim, aplicar PCA na matriz de atributos e avaliar o algoritmo mais representativo, determinando a estrutura de comunidades desejada. As próximas seções detalham as etapas da metodologia proposta.



**Figura 1. Metodologia para a descoberta de comunidades em traces de mobilidade reais**

#### 4.1. (Etapa 1) Conjuntos de Dados

Traces de mobilidade fornecem informações sobre a localização de dispositivos móveis no tempo. Analisando tais informações é possível obter características de mobilidade dos usuários móveis, incluindo a distância percorrida pelo usuário, a distância entre outros dispositivos, o tempo que eles passam juntos (i.e., tempo de contato), etc. Uma vez que pessoas se movem com um certo propósito (e.g. trabalho para casa), assumimos que suas localizações e características de mobilidade podem implicar em seus interesses e preferências. Por outro lado, assumimos também de forma análoga, que se usuários não compartilham características e locais em comum, eles tem menos chances de apresentarem relações sociais. Uma vez que tais características são encontradas, é possível agrupar indivíduos com comportamentos, preferências e interesses similares. Assim, pode-se obter mais sucesso na elaboração de modelos de mobilidades mais realísticos, aumentar a eficiência de algoritmos de encaminhamento de mensagens para redes oportunistas, no desenvolvimento de aplicações baseadas em contexto, entre outras aplicações. No entanto, a forma como os dados são tratados antes de alimentarem os algoritmos de agrupamento é de grande impacto no resultado final.

Os conjuntos de dados utilizados nesse trabalho foram escolhidos para abranger uma diversidade de cenários, considerando redes móveis humanas (GeoLife [Mani et al. 1999]) e veiculares (Taxis de São Francisco [Piorowski et al. 2009]). Ambos foram coletados usando coordenadas de GPS.

O trace GeoLife, é referente à mobilidade em vários cenários da cidade de Beijing, incluindo diferentes meios de transporte, caminhada, bicicleta e carro. Os traces representam trajetórias de GPS de 182 usuários, coletados durante um período de 3 anos, em instantes de 5 segundos. A trajetória do conjunto de dados é representada por uma sequência de pontos no tempo, cada uma contendo informação de latitude, longitude e altitude. O conjunto de dados contém 17.621 trajetórias com uma distância de 1.292.951 km e duração total de mais de 50.000 horas.

O trace de redes veiculares é referente aos traces de movimento de taxis na cidade de São Francisco/USA. O trace possui 483 usuários e foram coletados por 24 dias com amostras que variam de 1 a 3 minutos.

## 4.2. (Etapa 2) Pré-processamento dos Dados

Uma vez que estamos interessados na mobilidade dos usuários e em oportunidades de transmissão entre dispositivos, o conjunto de dados brutos é inadequado ao estudo apresentado nessa proposta e portanto requer pré-processamento. A primeira etapa consiste da seleção de uma região geográfica específica de cada conjunto de dados, onde existe maior concentração de usuários. Depois, as trajetórias dos usuários presentes nessa região, ainda em coordenadas geográficas em latitude/longitude, foram transformadas para o sistema de coordenadas cartesianas bidimensional - UTM. Para a construção da matriz de similaridades, foi necessário dividir a área geográfica em células quadradas de tamanho de 250 metros de lado [Nunes and Obraczka 2013]. Esse tamanho permite preservar as características espaciais dos usuários e o futuro tratamento dos atributos apresentados na seção seguinte. Vale comentar que diminuir o tamanho da célula, significa aumentar significativamente o número de atributos a serem processados posteriormente.

A *matriz de similaridades* contém as características geográficas e temporais dos usuários. A matriz proposta é constituída por  $N$  linhas (uma para cada usuário) e  $K + 2$  colunas, que representam: 1 para cada uma das  $K$  células, contendo o tempo de permanência de cada usuário em cada célula; 1 coluna para a velocidade média de cada nó móvel; e a última para o tempo médio de pausa dos mesmos. O tempo de pausa foi definido como o tempo transcorrido pelo nó quando sua movimentação é menor que 1 metro. O início de uma pausa é detectado quando um nó se desloca menos de um metro durante o tempo  $t_p$ , onde  $t_p$  é definido como a taxa de amostragem média do trace de mobilidade.

É importante salientar que a matriz de similaridades deve ser normalizada para que uma variável com valores mais altos que outras não domine as métricas de distância calculadas durante a análise dos agrupamentos. Além disso, propõe-se aplicar uma transformação não-linear à matriz de atributos, utilizando a função *logit* de verossimilhança logarítmica, antes da aplicação dos algoritmos de agrupamento. Quando aplicada aos dados normalizados esta transformação modifica as proporções das variáveis da matriz de atributos, para que os dados entre  $(0, 1)$  assumam valores reais, entre  $(-\infty, \infty)$ , e sejam simétricos em 0.5. Isso além de evidenciar as diferenças e semelhanças entre as observações para cada variável, também promove ganho real na detecção das similaridades durante o cálculo das métricas de distância, utilizadas pelos algoritmos de agrupamento.

Uma função *logit*() é definida como um logaritmo de probabilidades relativas. Se  $p$  é a probabilidade de um evento, então  $(1 - p)$  é a probabilidade de não observar o evento, e as probabilidades relativas do evento são  $p/(1 - p)$ . Logo, o *logit* de um número  $p$  entre 0 e 1 é dado por  $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ . Note que a *logit* não é definida quando  $p = 0$  ou  $p = 1$ . Uma solução para esse problema é adicionar algum valor pequeno,  $\epsilon$  ao numerador e ao denominador da função *logit*. Agora os dados estão prontos para a próxima etapa da metodologia que é a identificação de comunidades apresentada a seguir.

## 4.3. (Etapa 3) Definição das Comunidades

Após o pré-processamento dos dados, voltamos ao problema de encontrar a quantidade de comunidades existentes nos dados, para então aplicar os algoritmos apresentados na Seção 3. Para alcançar esses objetivos, lançamos mão da ferramenta matemática BIC que irá nos auxiliar nessa tomada de decisão. BIC é uma técnica utilizada extensivamente na literatura para encontrar o número mais adequado de agrupamentos  $k$  no

conjunto de dados, para o qual o valor do BIC é maximizado. Este valor  $k$  oferece o melhor tamanho para os agrupamentos de forma a representar a maior quantidade de observações possível, dentro de cada agrupamento, respeitando as métricas de similaridade.

Cada combinação do número de agrupamentos corresponde a modelos estatísticos diferentes, reduzindo o problema de encontrar o melhor número de agrupamentos a comparações entre um conjunto de modelos possíveis. Portanto, se vários modelos  $M_1, \dots, M_k$  são considerados, com probabilidades a priori  $p(M_k), \forall k = \{1, \dots, K\}$ , então pelo teorema de *Bayes*, as probabilidades a posteriori do modelo  $M_k$ , dadas pelo dado  $D$ , podem ser obtidas por [Chris Fraley 2002]:  $p(M_k|D) \propto p(D|M_k)p(M_k)$ .

Devido a forma de seleção do modelo (Bayesiano), se as probabilidades  $p(M_k)$  são iguais, então a escolha do modelo é dada pela maior verossimilhança (i.e., *maximum likelihood*) entre os modelos. Essa verossimilhança pode ser aproximada pelo BIC, de forma que:  $BIC_k = 2 \log p(D|\theta_k, M_k) - K \log(n) \propto 2 \log p(D|M_k)$ , onde  $K$  é o número de parâmetros independentes que devem ser estimados (e.g., o número de agrupamentos) para o modelo  $M_k$  e  $\theta_k$  é o vetor de parâmetros que maximiza a função de verossimilhança para o modelo  $M_k$ .

Após o cálculo do BIC para diversos tamanhos de grupos, decide-se pelo primeiro máximo local, que evidencia o número  $k$  de agrupamentos mais adequado ao modelo. Feito isso, os métodos de agrupamento apresentados na Seção 3 podem ser aplicados.

#### **4.4. (Etapa 4) - Aplicar os Algoritmos de Agrupamento**

Muitos métodos de agrupamento estão disponíveis na literatura, porém não há uma técnica universalmente aceita para encontrar todas as variedades de grupos e representar dados multidimensionais [Jain et al. 1999]. Assim, neste trabalho utilizamos 3 diferentes métodos de agrupamento apresentados na Seção 3, aplicados à dois conjuntos heterogêneos de dados reais de mobilidade humana e veicular, para identificar as similaridades entre os diversos indivíduos e agrupá-los adequadamente.

Além disso, uma vez que foram definidas as comunidades através dos diferentes métodos de agrupamento, se faz necessário decidir qual método foi capaz de capturar características de interesse que permitem identificar estruturas ou padrões nos traces de mobilidade, com o objetivo de caracterizar o comportamento da mobilidade humana. Nesse contexto, é interessante analisar as características dos componentes dentro de cada grupo, identificado através dos diferentes algoritmos de agrupamento, de forma a verificar se os mesmos foram capazes de agrupar as observações de acordo com as métricas de interesses definidas (e.g., tais como preferência geográfica, velocidade média e tempo médio de pausa). A próxima seção apresenta um método para realizar tal análise.

#### **4.5. (Etapa 5) - Redução de Dimensionalidade e Avaliação das Comunidades**

Com o objetivo de verificar se os agrupamentos foram capazes de distinguir indivíduos com preferências geográficas semelhantes no mesmo grupo e distintas em grupos diferentes, precisamos de um método que nos permita avaliar de forma consistente as diversas variáveis da matriz de atributos e identificar que parâmetros influenciam na atribuição das comunidades a cada um dos indivíduos. No entanto, tal matriz apresenta alta dimensionalidade (i.e., número de variáveis da ordem de  $10^2$ ). Assim, encontramos uma nova representação para os dados através da redução de dimensionalidade, através

da aplicação de PCA (análise de componentes principais) que nos permitiu: (1) explorar os dados de alta dimensionalidade e identificar padrões nos mesmos, como por exemplo, encontrar para que valores de um determinado componente principal os indivíduos estão agrupados. Exemplos aplicados serão apresentados nas análises de casos de uso na Seção 5; (2) visualizar os dados quando a dimensionalidade dos mesmos é reduzida para  $\mathbb{R}^3$  ou mesmo no  $\mathbb{R}^2$ ; (3) analisar os dados através do uso de ferramentas estatísticas, tais como densidade de probabilidade, clusterização, ou classificação.

Essa redução é realizada através da criação de novas variáveis que são funções (e.g., combinações lineares) das variáveis originais. Portanto, o método mapeia dados do espaço original (alta dimensão) para o espaço de baixa dimensão, enquanto mantém a informação de todas as variáveis disponíveis. A função principal da análise de componente principal (PCA), usada no presente trabalho, é reduzir a dimensionalidade dos dados e ao mesmo tempo contabilizar a maior variância possível do conjunto de dados originais. Assim, os dados são transformados em um novo conjunto de coordenadas ou variáveis, não correlacionadas, que são combinações lineares das variáveis originais. Dessa forma, o objetivo é ajudar no entendimento e identificação de padrões nos dados originais através da análise das observações obtidas no novo espaço. Aplicações da nossa análise serão apresentadas na próxima seção.

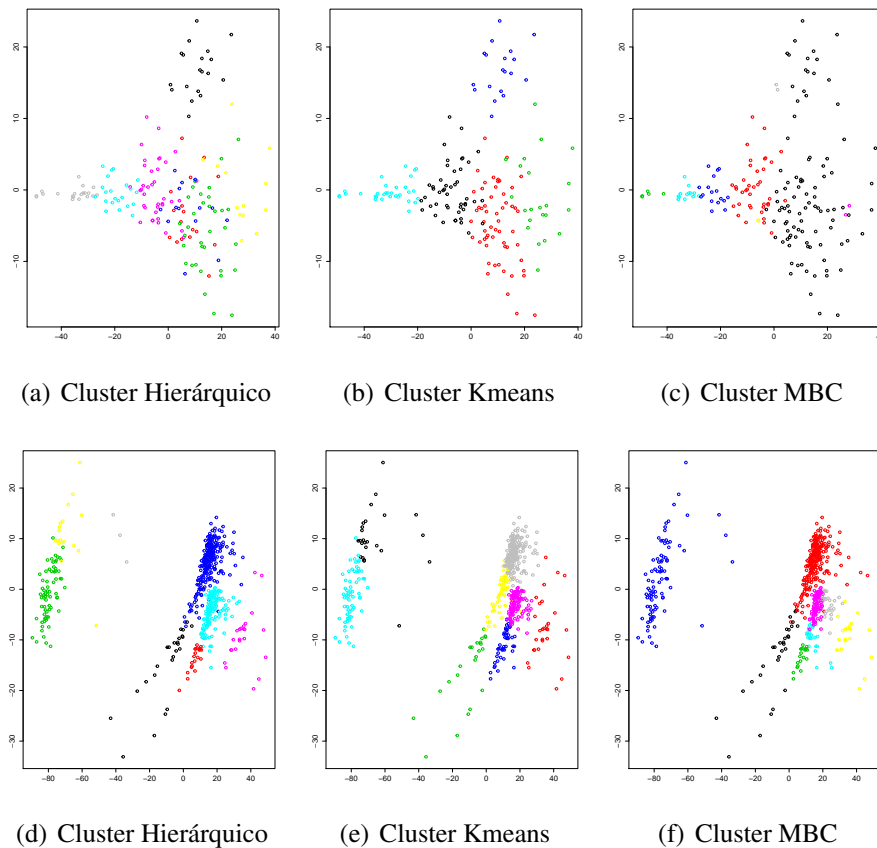
## 5. Estudos de Caso

Nessa seção a metodologia proposta é validada em dois estudos de caso correspondendo aos traces GeoLife e São Francisco, descritos na Seção 4.1. Mostraremos como a metodologia foi capaz de capturar os padrões de movimentação humana através da identificação das preferências geográficas dos usuários e da extração das estruturas de comunidade existentes nos traces. A metodologia proposta foi validada através da técnica PCA que permite extrair medidas estatísticas das estruturas de comunidade e compará-las. A técnica também permite uma identificação visual (plano 2-D) das preferências geográficas identificadas nas diferentes comunidades.

Cluster	1	2	3	4	5	6	7	8
HC - GeoLife	45	36	19	17	17	16	15	13
KMeans - GeoLife	40	28	23	19	18	17	17	16
MBC - GeoLife	97	39	17	13	5	3	2	2
HC - SF	240	81	77	69	25	23	18	3
KMeans - SF	186	97	67	56	54	29	24	23
MBC - SF	238	90	85	30	26	26	21	20

**Tabela 1. Tamanhos das comunidades para cada algoritmo de agrupamento aplicado aos traces de mobilidade.**

As estruturas de comunidades extraídas dos traces estudados foram obtidas através da aplicação do método de identificação de comunidades apresentado na Seção 4. O método identificou oito comunidades de tamanhos distintos, para ambos os traces, aplicando-se os 3 algoritmos de agrupamento descritos na Seção 3. Os resultados apresentados na Tabela 1 apresentam a quantidade de usuários identificada em cada uma das comunidades, para os diferentes algoritmos. Note que, a maioria das comunidades estão separadas em tamanhos similares, para os métodos de agrupamento k-means e hierárquico, fenômeno que já era esperado de acordo com as características desses algoritmos apresentadas na Seção 3. O método de agrupamento MBC, por outro lado, forma estruturas de comunidades maiores, devido ao seu comportamento na formação dos grupos.

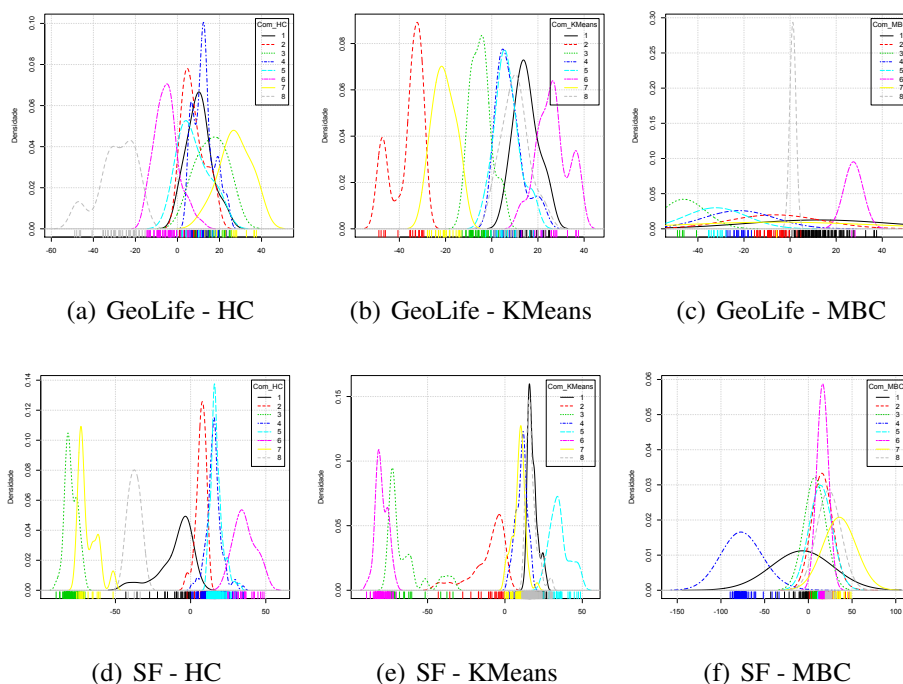


**Figura 2. Diferentes análises dos algoritmos de agrupamento estudados.**

Existe um equilíbrio entre a escolha do tamanho da comunidade e o algoritmo de agrupamento aplicado. O método de agrupamento MBC, utiliza como métrica para a formação de grupos a matriz de covariância, o que implica em grupos cujas observações dispõem-se no formato de uma elipsóide grande e alongada. Por outro lado, os métodos hierárquico e k-means geram grupos cujas observações estão dispostas de forma esférica e de volumes similares para ajustar os dados. Desta forma, mais de um grupo pode ser necessário (quando utilizando k-means e hierárquico) para aproximar a elipsóide encontrada pelo método MBC [Chris Fraley 2002]. Esse comportamento pode ser visualmente verificado, em ambos os traces, através da Figura 2, onde os algoritmos k-means e hierárquico apresentam grupos de forma arredondada (Figuras 2(a), 2(b), 2(d), 2(e)) e o algoritmo MD apresenta grupos maiores e com forma elíptica (Figuras 2(c) e 2(f)).

Além disso, a Figura 2 apresenta a visualização dos agrupamentos, para os traces GeoLife e São Francisco, onde cada ponto exibido nos gráficos corresponde a uma observação (i.e., um usuário no nosso caso). Utilizamos o PCA (*Principal Component Analysis*) para reduzir o número de dimensões dos dados e auxiliar na visualização. Os gráficos mostram o primeiro componente principal (PC1) no eixo y e o segundo componente principal (PC2) no eixo x. É possível observar nesta figura que através de apenas dois componentes principais, já é possível visualizar as comunidades de forma bem distinta e identificar intervalos em PC1 e PC2 onde cada comunidade reside.

Buscando verificar se o algoritmo de agrupamento foi capaz de capturar as prefe-

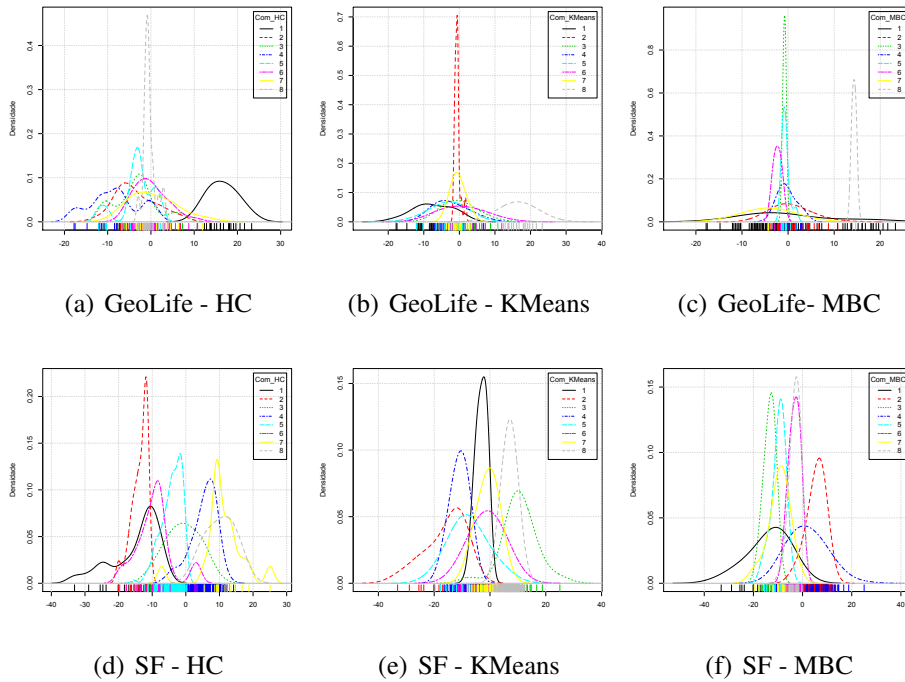


**Figura 3. Densidade das preferências geográficas (PC1) dos dados reduzidos utilizando PCA para três algoritmos de agrupamento, nos traces de mobilidade.**

rências geográficas e as características de mobilidade dos usuários nas diferentes comunidades, foi realizado um estudo das características apresentadas em cada comunidade. Inicialmente, as preferências geográficas dos usuários foram calculadas através do PCA. Os 126 atributos relativos ao tempo total que cada usuário passa em cada célula foram reduzidas para 2 componentes principais, que representam juntos 80% da variância dos dados. O primeiro componente principal responde por 60% dessa variância, oferecendo portanto uma boa representatividade das preferências geográficas dos usuários.

A Figura 3 apresenta a densidade das preferências geográficas obtidas através do PC1, para os três algoritmos de agrupamento, em ambos os traces de mobilidade estudados. É possível observar que o método de identificação de comunidades é capaz de diferenciar a predileção dos membros das diferentes comunidades por determinadas regiões geográficas. O método de visualização apresentado considera dois componentes principais, denominados PC1 e PC2, para exibir a preferência dos membros de cada comunidade por determinada área geográfica.

As mencionadas figuras também revelam que os algoritmos de agrupamento *k-means* e HC diferenciam as preferências geográficas dos usuários, bem como suas velocidades médias (Figura 5) para diferentes comunidades. Por exemplo, na Figura 3(b), *k-means* é capaz de diferenciar cinco diferentes regiões geográficas de interesse das comunidades 4, 5, 6, 7, 8. As comunidades 1,2 e 3, no entanto, não foram diferenciadas apenas com PC1, dado que suas curvas de densidade encontram-se muito sobrepostas. Considerando-se PC2, *k-means* é capaz de diferenciar as preferências geográficas das comunidades 1,2 e 3 (Figura 4(b)). Analogamente, observa-se que HC tem comportamento similar ao *k-means* e também é capaz de diferenciar as preferências geográficas dos usuários pertencentes as diferentes comunidades. Porém, MBC falha na representação de tais



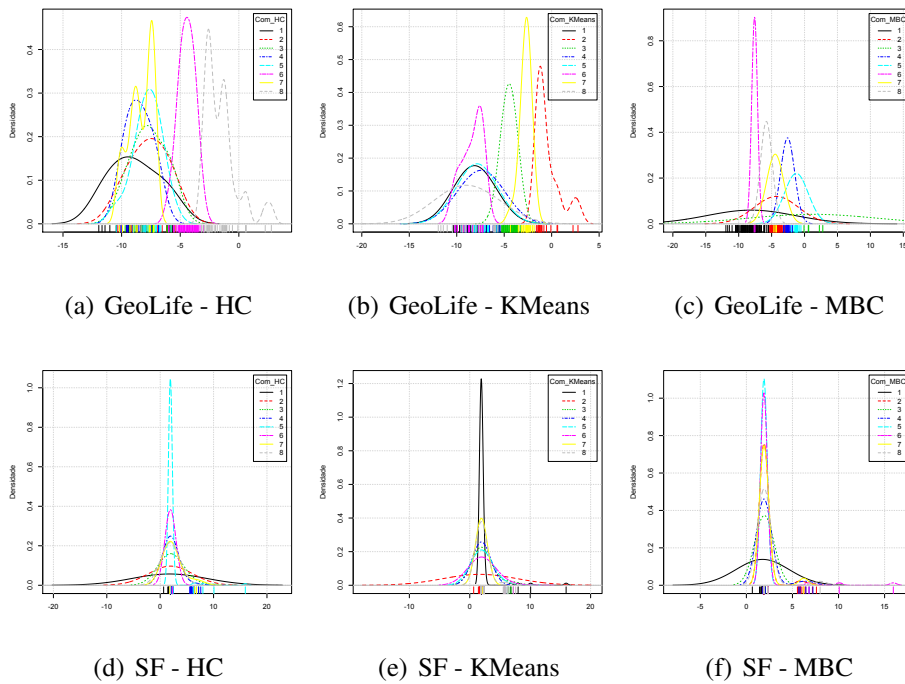
**Figura 4. Densidade das preferências geográficas (PC2) dos dados reduzidos utilizando PCA para três algoritmos de agrupamento, nos traces de mobilidade.**

preferências e na diferenciação das velocidades médias das diferentes comunidades. Esse comportamento esperado, pode ser justificado pelo comentado acima, onde MBC gera agrupamentos de tamanho maiores e de formato elipsoidal. Os comportamentos descritos podem ser estendidos para o trace de São Francisco, onde os algoritmos hierárquico e *k-means* melhor representam as preferências geográficas dos usuários.

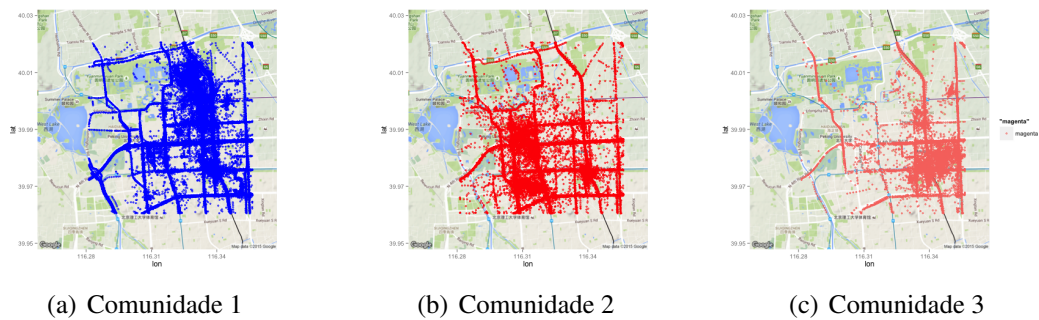
A validação do método para a extração das preferências geográficas das diferentes comunidades pode ser complementada através da Figura 6. Essa figura apresenta a localização ao longo do tempo dos usuários pertencentes as diferentes comunidades no mapa da cidade de Pequim, usando o método proposto com o algoritmo de agrupamento hierárquico. Aqui é possível verificar que de fato, os usuários de determinadas comunidades tem certas predileções por diferentes regiões geográficas.

## 6. Considerações Finais

O presente trabalho apresenta uma proposta de uma metodologia de detecção de estruturas de comunidades através do uso de registros reais de mobilidade. A metodologia também contempla a extração da preferência geográfica dos usuários móveis, através do uso de registros de mobilidade humana e/ou veicular, permitindo identificar diferentes padrões dependendo do universo temporal definido (e.g. diferentes dias da semana). A metodologia proposta é genérica o suficiente de forma que pode ser aplicada a um número grande de tipos de registros de mobilidade, ou seja, alta heterogeneidade dos dados. A metodologia foca em conjuntos de dados que podem ser facilmente coletados através de qualquer *smartphone* celular moderno, o que elimina a dependência de dados, muitas vezes difíceis de se obter, como os provenientes de operadoras de telefonia móvel, redes sociais online ou dados de homofilia. Apresentou-se também uma métrica de similaridade



**Figura 5. Densidade da velocidade média dos usuários, por comunidade, para os três algoritmos de agrupamento, nos traces GeoLife e São Francisco.**



**Figura 6. Preferência Geográfica de diferentes comunidades plotadas no mapa da cidade de Beijing, usando cluster hierárquico.**

dade baseado nas características intrínsecas do movimento, independente de agentes ou informações externas. Por fim, foi apresentado um estudo quantitativo comparando 3 métodos distintos de *clustering*, aplicados a 2 conjuntos de dados reais, um de movimentação humana e o outro veicular.

## Referências

- Chris Fraley, A. E. R. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.
- Chuah, M. and Coman, A. (2009). Identifying connectors and communities: Understanding their impacts on the performance of a dtn publish/subscribe system. In *ICCSE'09*, pages 1093–1098.
- Dasgupta, A. and Raftery, A. E. (1995). Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering. *Journal of the Americ. Stat. Association*, 93:294–302.

- Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification*. Wiley.
- Eagle, N. and (Sandy) Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput.*, 10(4):255–268.
- Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. In *of the 3rd ACM SIGSPATIAL*, pages 9–16, New York, USA.
- Giannotti, F., Nanni, M., Pinelli, F., and Pedreschi, D. (2007). Trajectory pattern mining. In *Proc of 13th ACM SIGKDD, KDD '07*, pages 330–339.
- Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.
- Han, B., Hui, P., Kumar, V. A., Marathe, M. V., Shao, J., and Srinivasan, A. (2012). Mobile data offloading through opportunistic communications and social participation. *Mobile Computing, IEEE Transactions on*, 11(5):821–834.
- Hui, P., Yoneki, E., Chan, S. Y., and Crowcroft, J. (2007). Distributed community detection in delay tolerant networks. In *2Nd ACM/IEEE MobiArch'07*, pages 7:1–7:8, New York, USA.
- Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review.
- Khetarpaul, S., Chauhan, R., Gupta, S. K., Subramaniam, L. V., and Nambiar, U. (2011). Mining gps data to determine interesting locations. In *Proc. of the 8th IIWeb'11*, pages 8:1–8:6.
- Li, F. and Wu, J. (2009). Localcom: A community-based epidemic forwarding scheme in disruption-tolerant networks. In *6th IEEE SECON'09.*, pages 1–9.
- Mani, D. R., Drew, J., Betz, A., and Datta, P. (1999). Statistics and data mining techniques for lifetime value modeling. In *Proc. of ACM SIGKDD'99*, pages 94–103.
- Motani, M., Srinivasan, V., and Nuggehalli, P. S. (2005). Peoplenet: Engineering a wireless virtual social network. In *Proc. of the 11th MobiCom'05*, pages 243–257.
- Newman, M. (2004). Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2):321–330.
- Nunes, B. A. A. and Obraczka, K. (2013). Modeling spatial node density in waypoint mobility. *IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems*, 0:453–457.
- Phithakkitnukoon, S., Smoreda, Z., and Olivier, P. (2012). Socio-geography of human mobility: a study using longitudinal mobile phone data. *PLoS ONE*, 7(6):e39253.
- Piorowski, M., Sarafijanovic-Djukic, N., and Grossglauser, M. (2009). CRAWDAD data set epfl/mobility (v. 2009-02-24). Downloaded from <http://crawdad.cs.dartmouth.edu/epfl/mobility>.
- Tang, L., Wang, X., and Liu, H. (2012). Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33.
- Zheng, Y., Xie, X., and Ma, W.-Y. (2009). Mining interesting locations and travel sequences from gps trajectories. In *ACM WWW 2009*. ACM WWW 2009.